Tesis de Maestría

Correlaciones periodístico-legislativos en la Argentina durante los años 2006-2011

Lenton, Antonio A.

2013

Universidad de Buenos Aires

Maestría en explotación de datos y descubrimiento del conocimiento

Tesis de Maestría

“Correlaciones periodístico-legislativos en Argentina en los años 2006 a 2011”

Presentada por:

Ing. Antonio A. Lenton

Dirigida por:

Dr. Carlos G. Diuk-Wasser

Buenos Aires, Argentina, 2013

Agradecimientos

A mi director, Greg

Por su asesoramiento para realizar esta tesis; por su entusiasmo para ponernos en marcha y su experiencia y guía para saber llegar.

A Mummy y Daddy

Por enseñarme que hacer la tarea es tan importante como saber hacerla.
Y por todo.

A Estelita y Miguel

Por aceptarme en la familia y brindarme su apoyo constante durante toda la maestría.

A Vale

Por ser mi luz, mi alegría, y mi reposo.

Tabla de Contenidos

Introducción

Las ciencias políticas cuantitativas aplican métodos estadísticos a datos legislativos y políticos en general, a fin de encontrar patrones tanto para describir como para intentar predecir el comportamiento de legisladores, jueces y distintos actores políticos. Existen múltiples resultados previos, como [Clinton 2004] que ubica a los legisladores en un espacio de posturas legislativas asignando un punto ideal o postura legislativa ideal a cada legislador. En [Martin 2002] se realiza un análisis dinámico de estos mismos puntos ideales y su evolución a lo largo del tiempo. Por otro lado [Sirovich 2003] analiza la entropía en las votaciones para determinar cómo afecta la fidelidad al partido al comportamiento de los legisladores.

Si bien estos resultados y análisis abundan en datos de Estados Unidos, éste es un análisis que no se ha encontrado aplicado a datos de la República Argentina, por más que las votaciones nominales legislativas a nivel nacional se encuentran publicadas rigurosamente a partir del 2006. Si bien en la Argentina y América Latina existen varios observatorios legislativos y sitios web[1] [2] que se dedican a catalogar y organizar la información disponible, son pocos los esfuerzos realizados hasta el momento en términos de utilizar data mining para poder extraer información relevante de estos datos. En [Alemán 2009] se realiza un análisis de los puntos ideales de legisladores en base tanto a las votaciones nominales como a los proyectos presentados en conjunto (co-sponsorship), pero tomando datos en Argentina solamente hasta el 2003, siendo que es justamente en el 2006 cuando, por un cambio del reglamento interno de la Cámara de Diputados, todas las votaciones legislativas pasan a ser nominales y a estar debidamente publicadas según proyecto y legislador.

Recientemente fue publicado un artículo en uno de los principales diarios de Argentina[3] que mostraba, en una sencilla tabla ordenada, la asistencia a las sesiones durante el 2011 de distintos senadores nacionales. Que este análisis trivial sea suficiente para publicar un artículo en un medio nacional parece indicar que hay una escasez de recursos disponibles para los analistas políticos y periodistas. Así mismo es notable cómo incluso el análisis más sencillo permite obtener información interesante que generalmente se ignora por más que los datos en bruto sean públicos.

En esta tesis se realizan varios análisis sobre los conjuntos de datos de las votaciones nominales de los diputados nacionales en Argentina durante los años 2006 a 2011, de los proyectos legislativos

1 http://americo.usal.es/oir/legislatina/base_de_datos.htm
2 http://www.adclegislativo.org.ar/
3 http://www.lanacion.com.ar/1554312-boletin-de-asistencias-al-senado-menem-fue-el-peor-alumno

presentados y de los artículos periodísticos publicados durante estos años. En particular, se determinan los puntos ideales de los legisladores en base a las votaciones nominales y se utilizan éstos como un indicador de la postura legislativa tanto de los legisladores como de los proyectos que presentan, se construye un clasificador para encontrar artículos periodísticos que mencionen determinados proyectos legislativos, y se muestra que la atención brindada por los medios a distintos proyectos varía según la postura legislativa de los proyectos. También se determina que la tasa de aprobación de los proyectos varía con la postura legislativa de los firmantes, o no, según la conformación de la cámara.

El trabajo se separa en tres partes, una por cada conjunto de datos. El orden en el que se presentará el trabajo es aproximadamente igual al orden en que fue realizado cronológicamente, analizando primero cada conjunto de datos por separado, comenzando por un recuento de cómo fueron adquiridos y limpiados los datos, para luego comenzar por distintos análisis exploratorios sencillos, encontrando resultados cada vez más elaborados. Sobre el final de cada parte se presentan los análisis que involucran los conjuntos de datos de partes anteriores.

La primera parte analiza más de 200000 votaciones nominales emitidas entre 2006 y 2011. Luego de obtener los puntos ideales se visualizan los perfiles de votación de los legisladores según el mismo punto ideal, se proponen varias técnicas para agrupar a los legisladores según sus votaciones, y se realiza un análisis de la fidelidad al partido observando la entropía de las votaciones dentro de cada bloque y entre bloques. Finalmente se visualizan las votaciones en 2 dimensiones utilizando MDS y PCA, como posible dirección para futuros análisis de mayor complejidad.

La parte de análisis de proyectos trabaja con más de 30000 proyectos legislativos presentados durante los años con los que se cuentan con datos de votaciones nominales. Comienza aplicando el punto ideal de los legisladores a los proyectos mismos, y luego observa cómo varía la tasa de aprobación de los proyectos según el punto ideal de los firmantes, en cada año. Esta parte es de menor extensión ya que se utiliza este conjunto de datos principalmente como vínculo entre los datos de las votaciones nominales mediante los puntos ideales de los legisladores, y los artículos periodísticos.

Por último la tercera parte extrae y analiza un corpus de más de 900000 artículos periodísticos de tres diarios nacionales (Clarín, Página/12 y La Nación) durante los mismos años para los que se cuenta con los datos legislativos. Esta parte nuevamente comienza con un trabajo fuerte de exploración y de limpieza de datos que brinda resultados interesantes en sí mismos. Se utilizan técnicas tales como Probabilistic Topic Modelling [Blei 2011] y Latent Semantic Analysis [Wiemer-Hastings 2004] para

analizar cuantitativamente el contenido de estos artículos y comparar los distintos medios analizados en base a su contenido. Se construye un clasificador que determina cuáles artículos periodísticos mencionan un proyecto de ley o la temática de la que habla el proyecto, y cruzando este resultado con los anteriores se obtiene la distinta mediaticidad que se le da a los proyectos de ley de distintas posturas legislativas, o la tasa de aprobación de proyectos según la cantidad de artículos que los mencionan.

Este trabajo no está fomentado por ninguna filiación política ni financiado por ningún partido, y se espera que los resultados sean de utilidad o al menos de interés para el público general, brindando algún sustento científico a las impresiones y sensaciones acumuladas en la vida política cotidiana. Si las ideas aquí generadas se pueden aplicar eventualmente en algún observatorio legislativo para mejor visualizar y comprender la información disponible hoy en día, enhorabuena.

Como los análisis de cada sección son mayormente independientes entre sí, el marco teórico necesario para cada caso se introduce al principio de cada capítulo, en vez de ser presentado todo junto en una sección.

Contexto histórico

Los años 2006 al 2011 presentan relativa estabilidad legislativa para la Argentina. Superada la crisis del 2001, el Partido Justicialista pasa a tener 129 bancas en la Cámara de Diputados en el año 2003, sobre un total de 257, alcanzando a tener la mayoría de las bancas por una diferencia mínima.

En la Argentina se realizan elecciones legislativas cada dos años, renovándose cada vez la mitad de las bancas de diputados. De esta manera durante el período analizado se realizan dos elecciones, en los años 2007 y 2009. Si bien las elecciones son en los años impares los legisladores electos asumen al inicio del año siguiente, por lo que los cambios en el comportamiento de la cámara debido a su nueva conformación son aparentes en los años pares.

En el 2005 el bloque justicialista se divide en el Frente para la Victoria con 115 bancas, y el bloque Peronista Federal con 29. Recién en el 2009 estos dos bloques dejan de controlar la mayoría de las bancas entre ambos y tienen que recurrir a buscar aliados entre otros partidos. Incluso entonces lo pueden hacer entre partidos bastante alineados como el Partido Nuevo, o el Justicialista Nacional.

El resto de la cámara, entre tanto, se encuentra dividida entre la Unión Cívica Radical con entre 25 y 40 bancas según el año, el ARI, PRO, Coalición Cívica y el Partido Socialista, con unas 10 bancas cada uno, y varios partidos menores. Los partidos ARI, Coalición Cívica, GEN y UPT forman una alianza

para las elecciones del 2007 alcanzando 18 bancas, pero para las elecciones del 2009 ya se presentan nuevamente por listas separadas.

Soporte digital

Todo el código desarrollado para esta tesis se encuentra disponible en el material de soporte digital adjunto a la presente tesis, en la carpeta `codigo`.

En todos los casos se utiliza el lenguaje de programación Python, o bien R. Python es un gran lenguaje de usos generales, flexible para cualquier tarea de extracción y procesamiento de datos, con librerías disponibles para cualquier tarea desde trabajar con una base de datos relacional hasta procesar html inválido.

En general se utiliza Python para las tareas de adquisición, extracción, limpieza y transformación de datos, y R para ajustar modelos y presentar los resultados en gráficos.

Para ver las versiones exactas de las dependencias utilizadas y dónde obtenerlas, ver el Apéndice A.

Análisis de votaciones nominales

Descripción, adquisición y limpieza de datos

Las votaciones de los diputados son nominales cuando se registra y publica el voto de cada diputado. Hasta el 2006 las votaciones también podían ser numéricas, en las que sólo se consignaban el total de votos alcanzados por la afirmativa o negativa. A principios del 2006 se modificó el reglamento interno de la Cámara de Diputados para que todas las votaciones fueran nominales, para mejorar la transparencia de los actos de gobierno.

Estos datos se encuentran disponibles de dos fuentes, ambos en el sitio oficial de la Dirección de Sistemas Electrónicos, dependiente de la Cámara de Diputados, y ambos en formato PDF:

- En las actas de votaciones por reunión[4], y
- en los informes detallados por legislador[5].

Como lo que interesa aquí es justamente el comportamiento de cada legislador se opta por descargar y trabajar con los informes detallados por legislador.

Se realiza primero un contacto infructuoso con la Dirección de Sistemas Electrónicos para ver si es posible acceder a la base de datos que contiene la información, o bien a los mismos datos en un formato más estructurado, pero lamentablemente la información sólo se encuentra disponible al público tal y como se encuentra en estos reportes disponibles para descargar en la página web de la Cámara de Diputados.

Se construye entonces un script Python que descarga automáticamente los informes de todos los diputados de los años en cuestión, y extrae la información de las votaciones. El código para obtenerlo se encuentra en el archivo `pull_votes.py` y utiliza la misma técnica que se utiliza para obtener todos los conjuntos de datos de esta tesis: partiendo de una URL inicial se determina cuáles links en esa página corresponden a pdfs de votaciones, y se busca y procesa cada una en orden. El procesado de los pdfs es relativamente similar al del html, con un paso adicional previo que convierte el pdf a un xml específico. Para esta conversión se utilizó la librería `pdfminer`[6] de Python.

4 Por ejemplo http://www1.hcdn.gov.ar/dependencias/dselectronicos/actas/2010/128OE18_01_R28.pdf
5 Por ejemplo http://www1.hcdn.gov.ar/dependencias/dselectronicos/actas/Informe 2010/ALFONSIN, Ricardo Luis.pdf
6 http://www.unixuser.org/~euske/python/pdfminer/

La calidad de este conjunto de datos es notablemente bueno, ya que los mismos se encuentran cuidadosamente cargados y no se encuentran errores tipográficos ni datos faltantes. En su momento únicamente hubo un archivo faltante, los votos de Oscar Ernesto Rodríguez del 2006, que por alguna razón el enlace para descargar ese archivo estaba roto. Los primeros resultados son elaborados sin considerar este diputado, y desde entonces el problema fue solucionado en el sitio web.

Así se obtienen los datos de votaciones que se resumen en la Tabla 1.

Año	N° de Votaciones	N° de Votos emitidos
2006	321	80812
2007	206	52271
2008	123	31362
2009	119	30326
2010	64	15803
2011	55	13904

Tabla 1: Votos extraídos por año

Un primer rasgo notable de los datos es la disminución de la cantidad de votaciones progresivamente en cada año. Esto no es un artefacto de la extracción de los datos, efectivamente sucedió así: en los últimos años analizados hay una menor cantidad de votaciones debido a la falta de quorum para realizar la votación. Tanto el oficialismo como la oposición a veces optan por indicar a sus legisladores que no bajen a la cámara para una votación, cuando parece que la votación no va a ser favorable. Los últimos años son más trabados, por lo que muy seguido uno u otro lado utiliza este recurso. En estos casos al no realizarse la votación no se extrae ningún dato en los informes de votaciones de los legisladores. Si bien se podría realizar algún análisis con la asistencia de la reunión y el orden del día, no es motivo de estudio en el presente trabajo.

En cada votación cada diputado tiene cuatro alternativas: votar por Afirmativo, votar por Negativo, votar por Abstención o estar Ausente. En la Tabla 2 se muestran brevemente los porcentajes obtenidos de cada alternativa, por año.

Año	Afirmativo	Negativo	Ausente	Abstención
2006	56.34%	7.48%	34.09%	2.09%
2007	49.74%	7.69%	39.97%	2.60%
2008	59.79%	14.03%	24.29%	1.88%
2009	55.25%	8.74%	33.87%	2.15%
2010	60.31%	20.02%	16.47%	3.20%
2011	61.64%	9.39%	27.59%	1.37%

Tabla 2: Porcentajes de votos de cada tipo, por año

Resulta curioso que la mayoría de los años el Ausente tiene un mayor porcentaje que el voto Negativo. Siendo que algunas veces los legisladores utilizan el Ausente como recurso para evitar que una votación se lleve a cabo, se considera la posibilidad de tomar al Ausente como similar al voto Negativo. Como habría que tener cuidado de sólo imputar como “similar al negativo” aquellos ausentes que no tienen otro justificativo como una licencia oficial o falta prolongada por viaje o enfermedad, se opta por simplemente interpretar a los Ausentes como un dato faltante en aquellos algoritmos que lo permiten, o como un valor intermedio entre Afirmativo y Negativo, junto con la Abstención.

Análisis de los puntos ideales de los legisladores

Los estudios sobre comportamiento legislativo intentan descubrir las preferencias de los legisladores, analizando los datos de votaciones nominales. Se supone que cada proyecto de ley corresponde a un punto en un espacio de baja dimensionalidad, y cada legislador tiene su postura legislativa “ideal” en este espacio. La utilidad de un proyecto para cada legislador decae con la distancia entre su punto ideal y el proyecto. En [Clinton 2004] se muestra un modelo bayesiano y un método de inferencia basado en Cadenas de Markov – MonteCarlo (MCMC, por sus siglas en inglés) para estimar estos puntos ideales. Este método es flexible y se puede aplicar a distintas variantes y métodos de votación, además de proveer mejor discriminación para posturas legislativas extremas que otras alternativas tradicionalmente empleadas como el análisis de componentes principales [Clinton 2004].

El atractivo de estimar estos puntos ideales de los legisladores es que hacen que el comportamiento legislativo sea susceptible de un análisis cuantitativo, haciendo que el estudio de la política legislativa tenga un sustento cuantitativo. Nos permite describir a los legisladores, caracterizándolos sin tener que conocer su fondo ideológico ni su historia.

Se han utilizado para el estudio del congreso de los Estados Unidos [Canes-Wrone 2002], de las legislaturas de distintos estados [Wright 2002], y las votaciones de la Corte Suprema [Martin 2002].

En el modelo bayesiano que se describe en [Clinton 2004], los datos constan de *n* legisladores que participan de *m* votaciones separadas. En la votación de cada proyecto *j*=1...*m* los legisladores pueden optar entre un "Voto Afirmativo" o un "Voto Negativo". Se asume que cada dato $y_{ij} = 1$ si el legislador *i* vota Afirmativo en la votación *j*, y que $y_{ij} = 0$ si vota Negativo, sin pérdida de generalidad.

Cada votación j se representa en el modelo con dos parámetros ζ_j y ψ_j, ambas posiciones en un espacio legislativo de *d* dimensiones. Para este trabajo se toma $d = 1$. El punto ζ_j es la posición Afirmativa del proyecto, mientras que ψ_j es la posición Negativa. Los legisladores que tengan un punto ideal más próximo a ζ_j votarán más probablemente por el Afirmativo, mientras que es más probable que los legisladores con un punto ideal más próximo a ψ_j voten por Negativo.

Se asume que los legisladores tienen una función de utilidad cuadrática sobre el espacio legislativo:

$$\begin{aligned} U_i(\zeta_j) &= -\|x_i-\zeta_j\|^2+\eta_{ij} \\ U_i(\psi_j) &= -\|x_i-\psi_j\|^2+\nu_{ij} \end{aligned}$$

Aquí x_i es el punto ideal del legislador *i*, η_{ij} y ν_{ij} son errores o elementos estocásticos de la utilidad, y la doble barra indica la distancia euclídea. Para que se cumpla la suposición de que cada legislador maximiza su utilidad, en los datos debe cumplirse entonces que:

$$y_{ij}=\begin{cases} 1 & si \quad U_i(\zeta_j)>U_i(\psi_j) \\ 0 & si \quad U_i(\zeta_j)\le U_i(\psi_j) \end{cases}$$

Se asume que η_{ij} y ν_{ij} siguen una distribución normal con $E(\eta_{ij}) = E(\nu_{ij})$, que $\text{var}(\eta_{ij} - \nu_{ij}) = \sigma^2$, y que los errores son independientes tanto entre legisladores como entre votaciones.

De aquí que la probabilidad de que un legislador emita un voto Afirmativo es:

$$\begin{aligned} P(y_{ij}=1) &= P(U_i(\zeta_j) > U_i(\psi_j)) \\ &= P(\nu_{ij}-\eta_{ij} < \|x_i-\psi_j\|^2-\|x_i-\zeta_j\|^2) \\ &= P(\nu_{ij}-\eta_{ij} < 2(\zeta_j-\psi_j)x_i+\psi^2-\zeta^2) \\ &= \Phi(\alpha_j x_i-\beta_j) \end{aligned}$$

donde $\alpha_j = 2(\zeta_j - \psi_j)/\sigma_j$, $\beta_j = (\zeta_j^2 - \psi_j^2)/\sigma_j$, y Φ denota la función de distribución normal estándar.

En el caso de un espacio de votaciones unidimensional resulta entonces que si se cuenta con *n*

legisladores y m votaciones nominales, resulta un modelo estadístico con $n + 2m$ parámetros.

El problema con este modelo es que todo salvo las votaciones de los legisladores es una variable oculta: tanto los puntos ideales de los legisladores, los parámetros α_j, β_j, y las utilidades obtenidas por los legisladores. Si se pudieran imputar valores para los parámetros de las votaciones y las utilidades obtenidas, entonces los puntos ideales se podrían calcular mediante una regresión. De la misma manera, si se pudieran imputar valores para los puntos ideales y utilidades obtenidas, los parámetros de las votaciones también podrían ser estimados mediante regresión.

El algoritmo de MCMC realiza estas imputaciones y regresiones repetidas veces, comenzando con valores arbitrarios y alternando entre simulación de los puntos ideales, parámetros y utilidades.

Se utiliza el paquete `pscl` de R para calcular los puntos ideales de los datos obtenidos. Este paquete, desarrollado por el Laboratorio de Ciencias Políticas Computacionales de la universidad de Stanford, implementa el algoritmo que se describe en [Clinton 2004].

El código utilizado para correr el algoritmo se encuentra en el archivo `ideal_points.R` y generó los datos de los archivos `ideal_points_*.csv`. El algoritmo permite informar votos de abstención e incluso los ausentes, por lo que no se descarta ningún dato extraído.

Luego de obtener los puntos ideales se los grafica en orden creciente separados por año, como se muestra en la Figura 1.

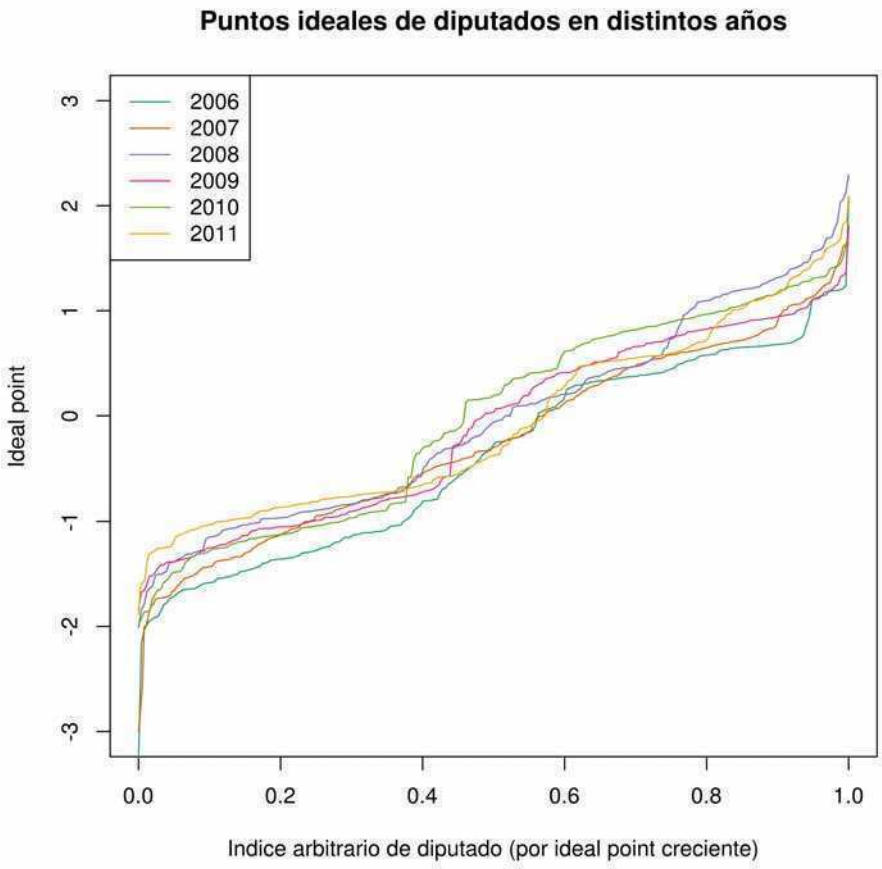

Figura 1: Puntos ideales de los legisladores, según el año

En esta Figura se muestra una curva por año. En el eje X se ubican todos los legisladores ordenados por punto ideal creciente, y se le asigna la abscisa 0 al legislador con el menor punto ideal, y la abscisa 1 al legislador con el mayor punto ideal de cada año. En el eje de las ordenadas se muestra el punto ideal correspondiente de cada legislador.

En primera instancia los resultados no se encuentran normalizados, así que si bien el rango de valores obtenido es similar para todos los años, no se pueden sacar conclusiones entre los valores del punto ideal de un legislador entre años distintos. En particular, no se puede decir que dos legisladores con un mismo punto ideal en años distintos tengan la misma postura legislativa. Ver la sección de *Normalización de los puntos ideales y evolución histórica*.

La Figura 1 parece indicar una menor cantidad de legisladores con un punto ideal cercano a cero. Antes de darle una interpretación a este fenómeno, y para visualizar esto mejor, se elaboran

histogramas de los legisladores, según su punto ideal. Estos histogramas se muestran en la Figura 2. Efectivamente, en cada año se observa una naturaleza bimodal en los resultados, con una agrupación de legisladores cercano al punto ideal -1, y otro grupo con un punto ideal cercano al +1.

Utilizando la librería `mixtools` de R se ajusta una mezcla de dos normales a los datos, y las dos curvas normales obtenidas para cada año se muestran superpuestas en los histogramas de la Figura 2. Esta librería utiliza un algoritmo de esperanza–maximización para buscar estimadores de máxima verosimilitud para los datos obtenidos [Benaglia 2009].

Es notable que en los datos del 2008, a diferencia de los otros años, parece haber un número de legisladores con un punto ideal cercano al 0. Cabe mencionar que con esto no se está intentando proponer el modelo que mejor se ajuste a los datos, sino sólo visualizar una tendencia general detectada. De hecho, si se calcula el criterio de información bayesiano (BIC, por sus siglas en inglés) como medida de bondad de ajuste para modelos que consisten de mezcla de 2, 3, 4 y 5 normales, de media y varianza variable, se ve que en todos los años se ajusta mejor una mezcla de entre 3 y 5 normales. La Tabla 3 muestra estos valores, resaltando el modelo de mejor BIC para cada año. El código para obtener estos valores se encuentra en `ideal_points/fitted_histograms.R`.

Cantidad de componentes	2006	2007	2008	2009	2010	2011
2	627.8606	1025.5564	667.0530	585.3996	631.5615	804.4518
3	618.4118	1019.6108	666.3836	583.0282	613.8636	772.8996
4	573.7715	1037.3991	670.1754	582.6723	611.4650	780.8826
5	588.5343	1038.0667	667.5839	591.2126	607.8682	783.1208

Tabla 3: BIC para mezclas de distintas cantidad de componentes normales, según el año

Volviendo a la interpretación de los datos en Argentina, en estudios como [Clinton 2004] aplicados a datos de los Estados Unidos, la dimensión sobre la que se ubican los legisladores se divide en Demócratas y Republicanos, y en otros contextos [Rosenthal 2004] se sabe interpretar como la izquierda y la derecha política. Cuando se ubican los legisladores en un espacio de dos dimensiones, puede interpretarse como una dimensión de temas económicos y otra de temas sociales, por ejemplo. Para asignar esta interpretación a los datos de Argentina, se separan los datos por bloque. En la Figura 3 se muestra un boxplot por partido, para todos los partidos con más de un legislador.

La Figura 3 permite ver adónde se agrupan en general los puntos ideales de los legisladores de cada bloque. Aquí se ve cómo se ubican los puntos ideales de los legisladores del Frente para la Victoria y el oficialismo de un lado, en este caso los puntos ideales menores a cero, y la UCR, ARI y el resto de la oposición en los puntos ideales mayores a cero.

Resulta interesante ver en este gráfico que no se trata de una dimensión de autodenominada izquierda contra derecha política, ya que el Partido Socialista se encuentra en el centro del espectro y el ARI (partido de concepción socioliberal de centro-izquierda) se ubica del mismo lado y relativamente próximo a Acción por la República (partido más de fondo liberal-conservador de derecha); se trata de resultados basados estrictamente en el comportamiento de los legisladores en las votaciones.

La interpretación de estos puntos ideales en la presente tesis es que se pueden tomar como un "índice de oficialismo" del legislador, donde los legisladores oficialistas son los que muestran un punto ideal menor a cero, y los legisladores de oposición son los de un punto ideal mayor que cero.

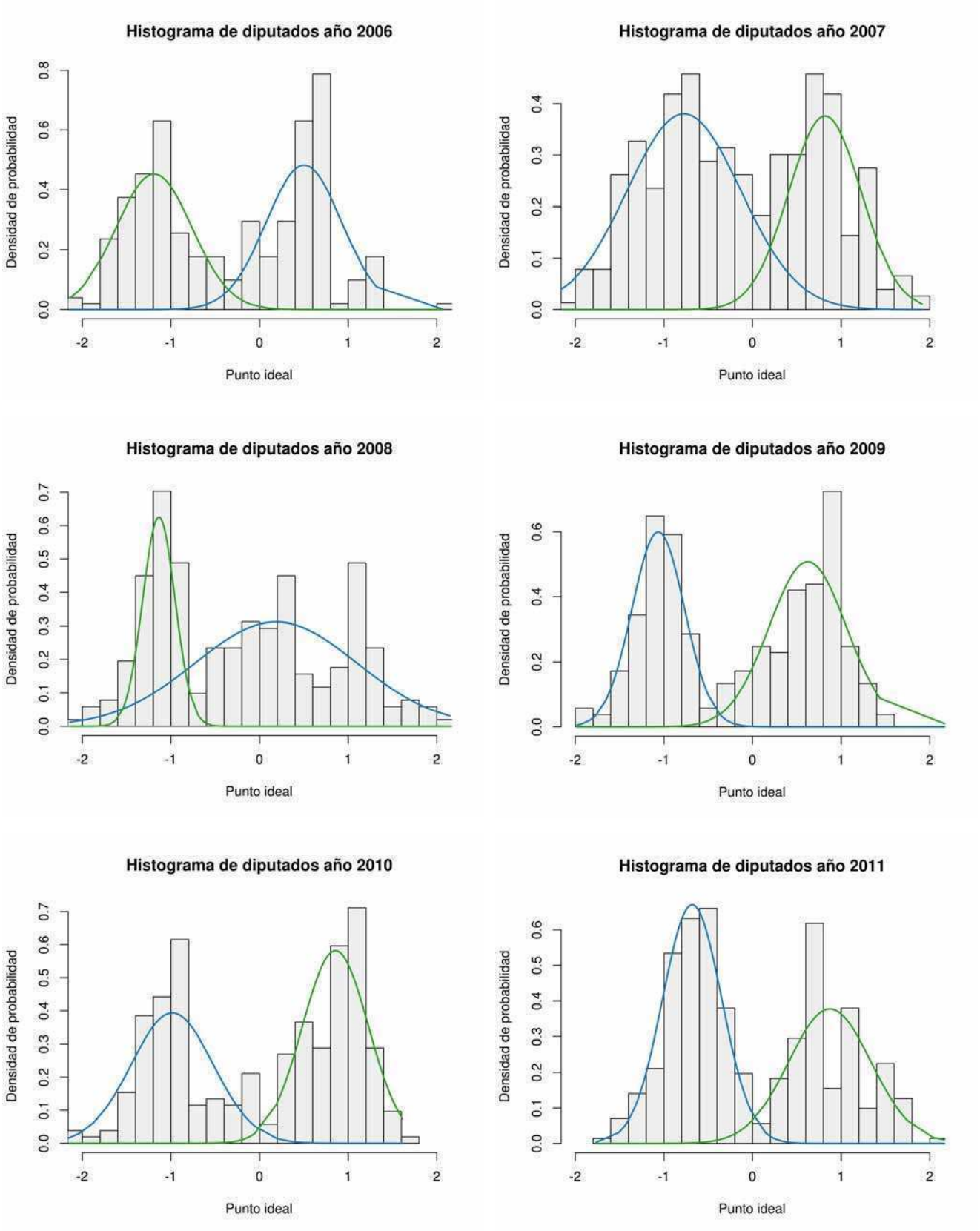

Figura 2: Histograma de puntos ideales, por año

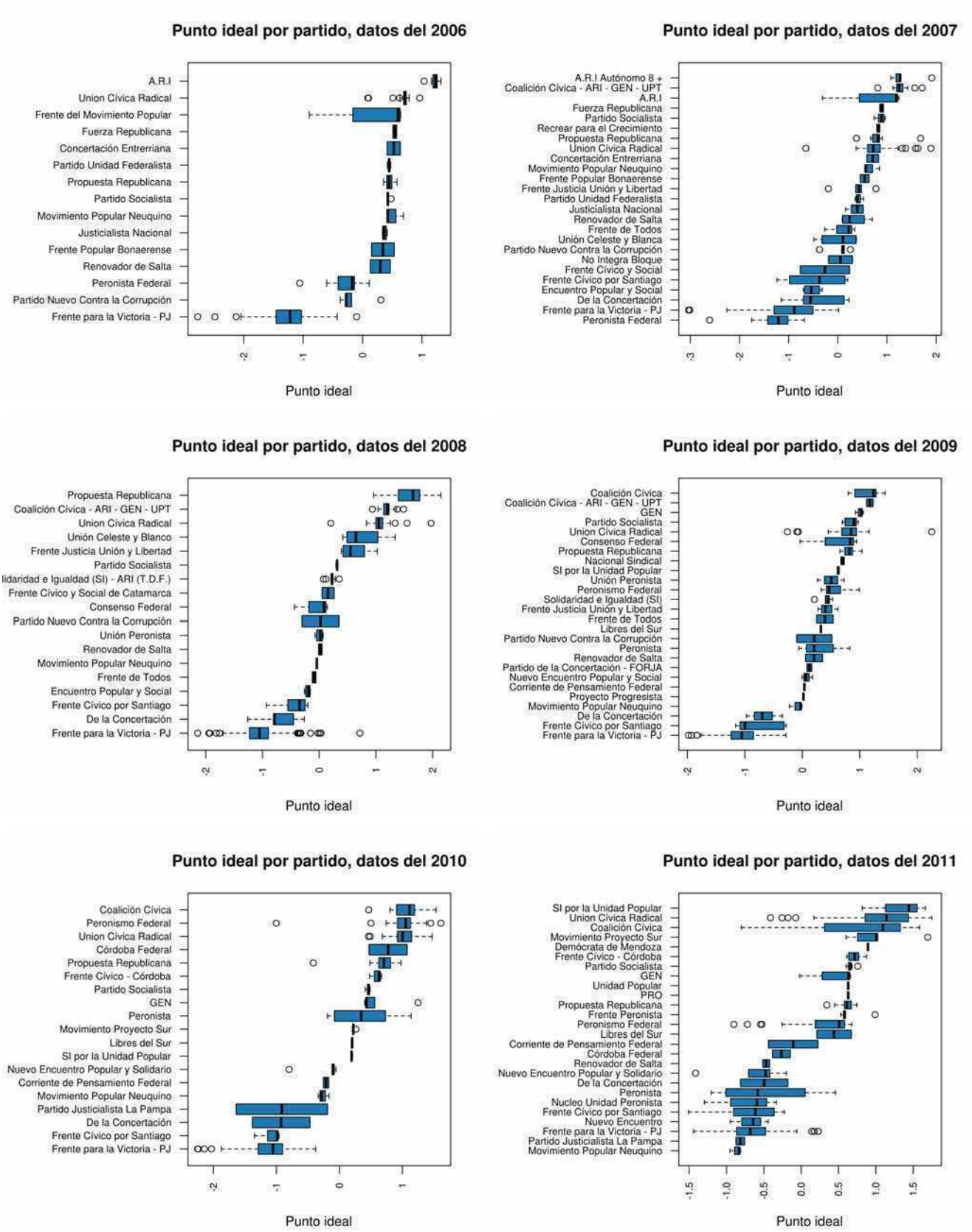

Figura 3: Boxplots de puntos ideales por partido

Volviendo brevemente a los datos del 2008 que en la Figura 2 parecen tener tres picos en vez de dos, en la Figura 3 ahora se puede apreciar que estos legisladores centrales son de partidos como el Consenso Federal, la Unión Peronista o el Partido Nuevo, que en el 2008 se alternaban en una postura de oficialismo y oposición.

Más allá del caso del 2008, en los datos obtenidos son los bloques chicos los que terminan llenando el espacio en el centro del espectro, teniendo que negociar posturas y proyectos con uno u otro de los partidos mayoritarios.

Para ver que los bloques grandes son los que establecen puntos ideales marcadamente distintos, se grafican los puntos ideales de cada año, removiendo los bloques pequeños.

En la Figura 4 se muestran los puntos ideales legislador por legislador, pero mostrando sólo los partidos con mayor número de bancas. Para cada año sólo se muestran los 8 partidos con mayor número de bancas, o menos en caso de empate. No se ordenan los legisladores por punto ideal creciente para apreciar mejor la cantidad de diputados en cada bloque, pero en cada año se puede apreciar la brecha que existe en posturas legislativas de los partidos mayoritarios de oficialismo y oposición.

Comportamiento legislativo según punto ideal

Una vez obtenidos los puntos ideales de los legisladores todo el análisis cobra una nueva dimensión de interés. Por ejemplo, se pueden tomar los porcentajes de votos de cada tipo como se muestra en la Tabla 2, y graficar según el punto ideal. En la Figura 5 se tomaron los legisladores agrupados en 25 cuantiles para cada año, es decir que cada punto representa un promedio sobre aproximadamente 10 legisladores, y se muestra el porcentaje de votos de cada tipo para cada año.

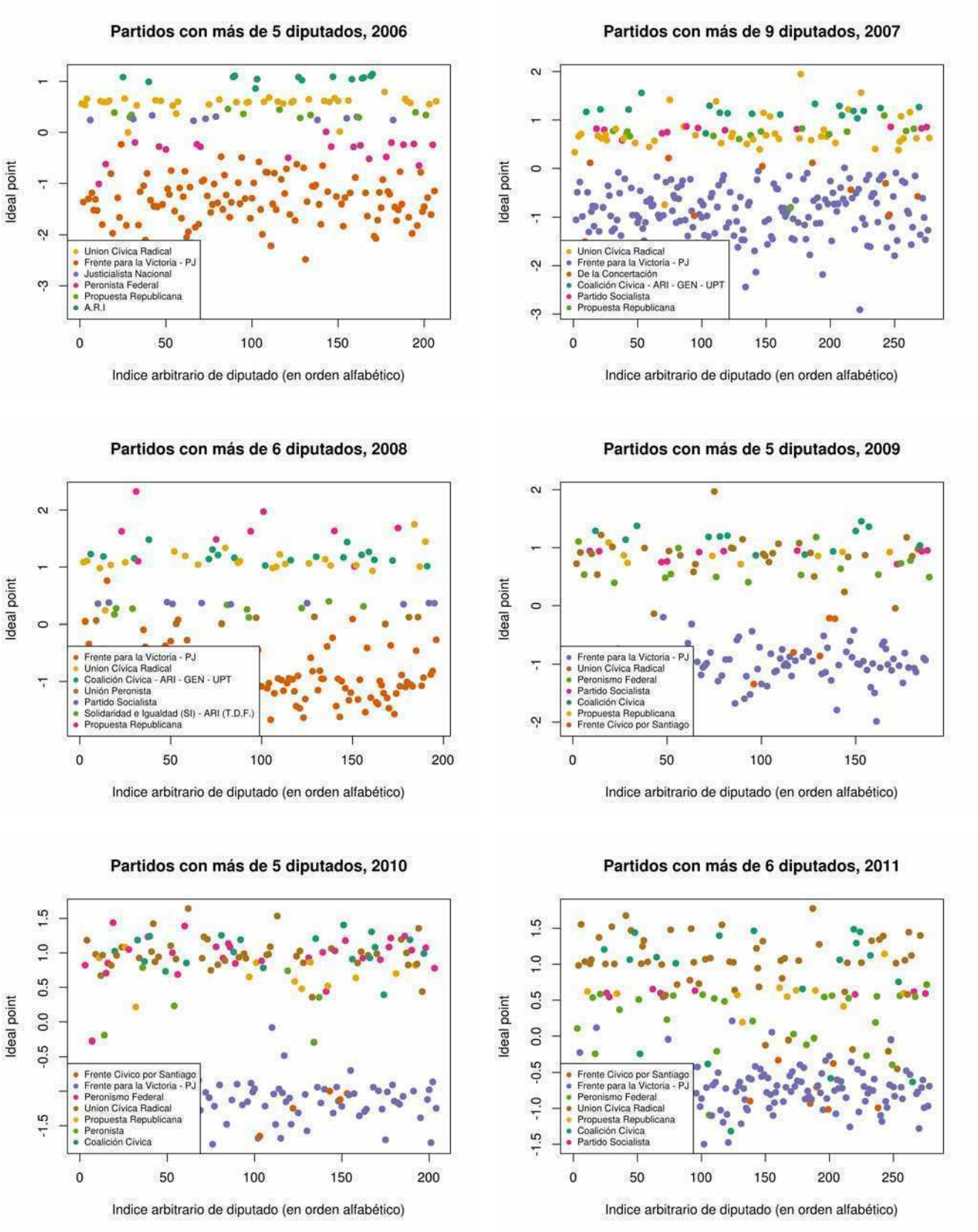

Figura 4: Puntos ideales por legislador, para partidos mayoritarios

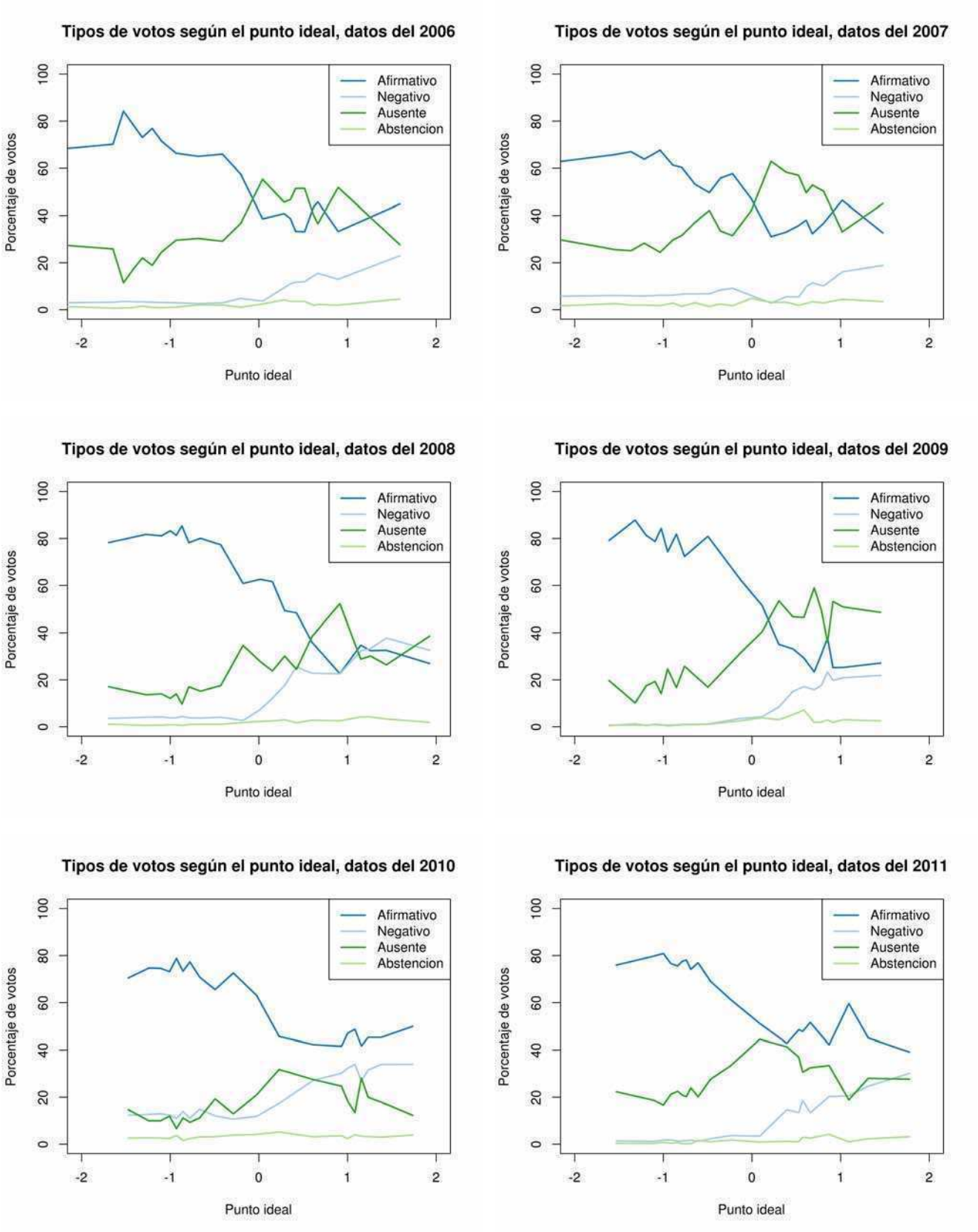

Figura 5: Porcentaje de votos de cada tipo según punto ideal, por año

Aquí se puede apreciar el distinto comportamiento de los legisladores según el punto ideal. Los legisladores oficialistas tienen un mayor porcentaje de votos por el Afirmativo. Los de punto ideal opositor recurren más a ausentarse, y tienen notablemente más votos por el Negativo. Éste no es un patrón incidental de un año, el mismo perfil se repite durante todos los años analizados.

Intervalos de confianza de los puntos ideales

Para tener una idea de qué tan robusto es el método utilizado, y cuánta precisión se le puede asignar a los puntos ideales obtenidos, se calculan los intervalos de confianza del 95% de estos valores.

El mismo algoritmo de MCMC que estima los puntos ideales nos brinda intervalos de confianza para cada punto ideal calculado. Los mismos se desprenden de la manera en que se ejecuta el algoritmo, realizando sucesivas iteraciones calculando posibles valores para los puntos ideales. Luego de 10000 iteraciones, el intervalo de confianza se puede estimar como el intervalo más chico que cubre el 95% de los valores calculados en cada paso de la traza. Esto se muestra en la Figura 6.

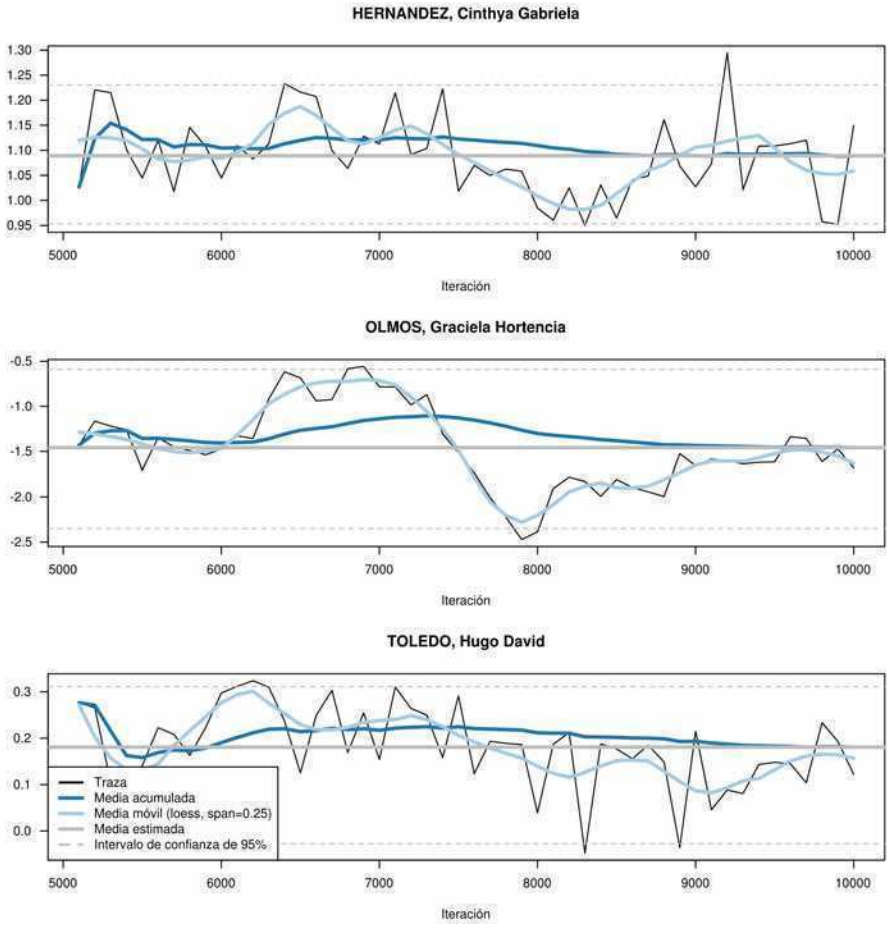

Figura 6: Trazas de ejecución de MCMC para tres legisladores

De esta manera se construye un gráfico que muestra el intervalo de confianza de 95% y en la Figura 7 se muestra para los legisladores del 2006. En este gráfico se muestra un punto por cada legislador, ordenados por punto ideal estimado, aunque por razones de espacio no se incluyen los nombres del legislador para todos los puntos.

El código para generar este gráfico se encuentra en el archivo `ideal_points/plot_ideal_ic.R`. La función `plotIC` es una versión modificada de la función `plot.ideal` brindada por la librería `pscl`.

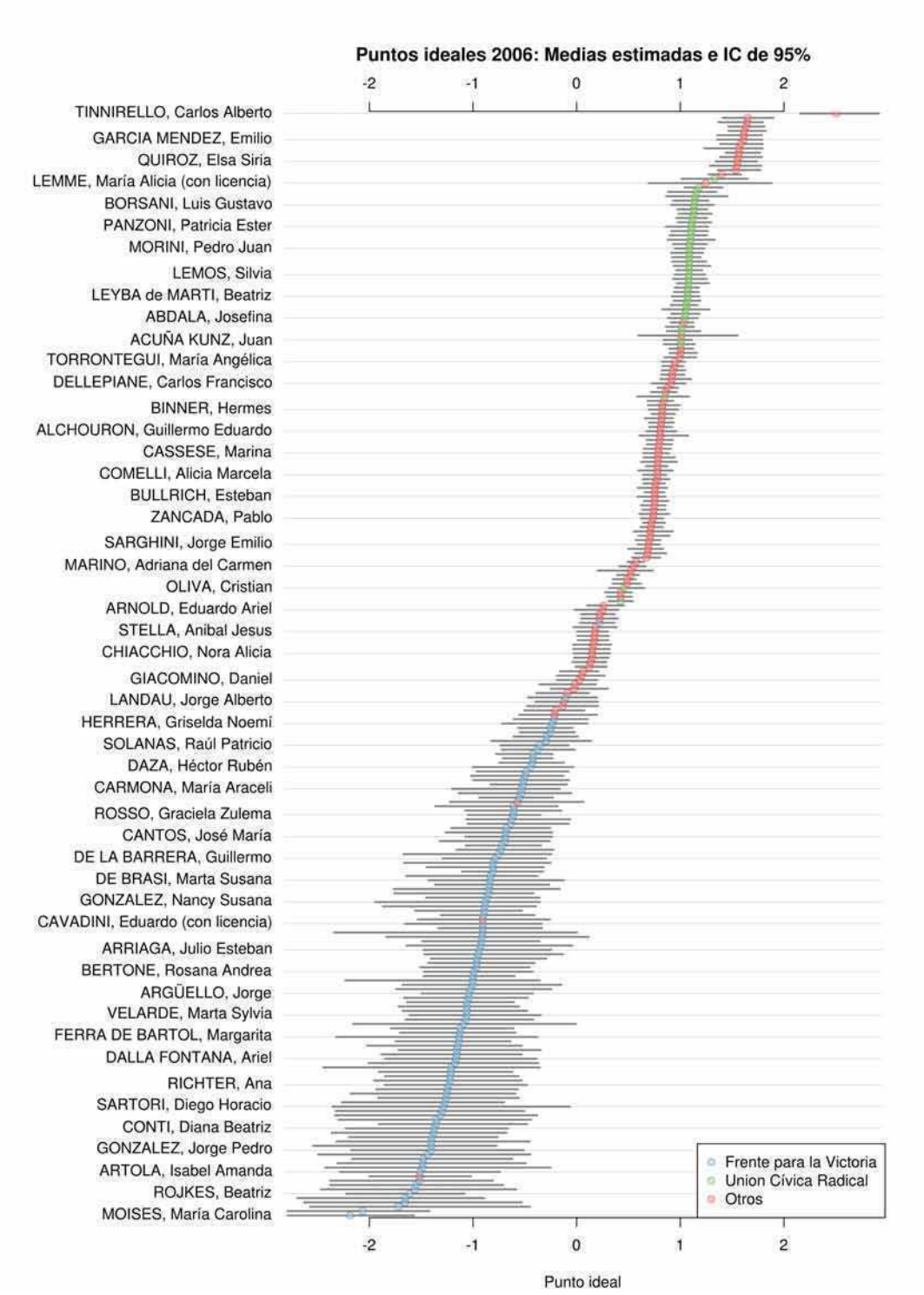

Figura 7: Puntos ideales del 2006, con intervalos de confianza de 95%

Como se observa en la Figura 7, las estimaciones del algoritmo no permiten sacar conclusiones sobre legisladores con puntos ideales muy similares, pero en general debería poder distinguirse claramente el

comportamiento de legisladores con puntos ideales de signos opuestos.

Un caso notable es el de Ricardo Wilder, diputado por Tierra del Fuego por el Frente de Unidad Provincial. Es el punto de la clase “Otros” en rojo de más abajo y a la izquierda, en el medio del bloque del Frente para la Victoria, con un punto ideal de -1.72. En el 2006 se perfilaba como un aliado permanente del oficialismo, según reportaron los medios en ese momento[7], y aparentemente terminó siendo así en los votos. Este caso es notable porque el algoritmo permite descubrir estas alianzas sin estar informado más que de las votaciones efectivas de los legisladores.

Resultados similares se obtuvieron para los restantes años. No se incluyen los gráficos de los años restantes, salvo por el del 2011, que se considera interesante por contar con menos votaciones ese año, por lo que intuitivamente tendría intervalos de confianza más grandes. Se muestran los intervalos de confianza para este año en la Figura 8.

7 http://www.lanacion.com.ar/768710-el-oficialismo-busca-diputados-aliados

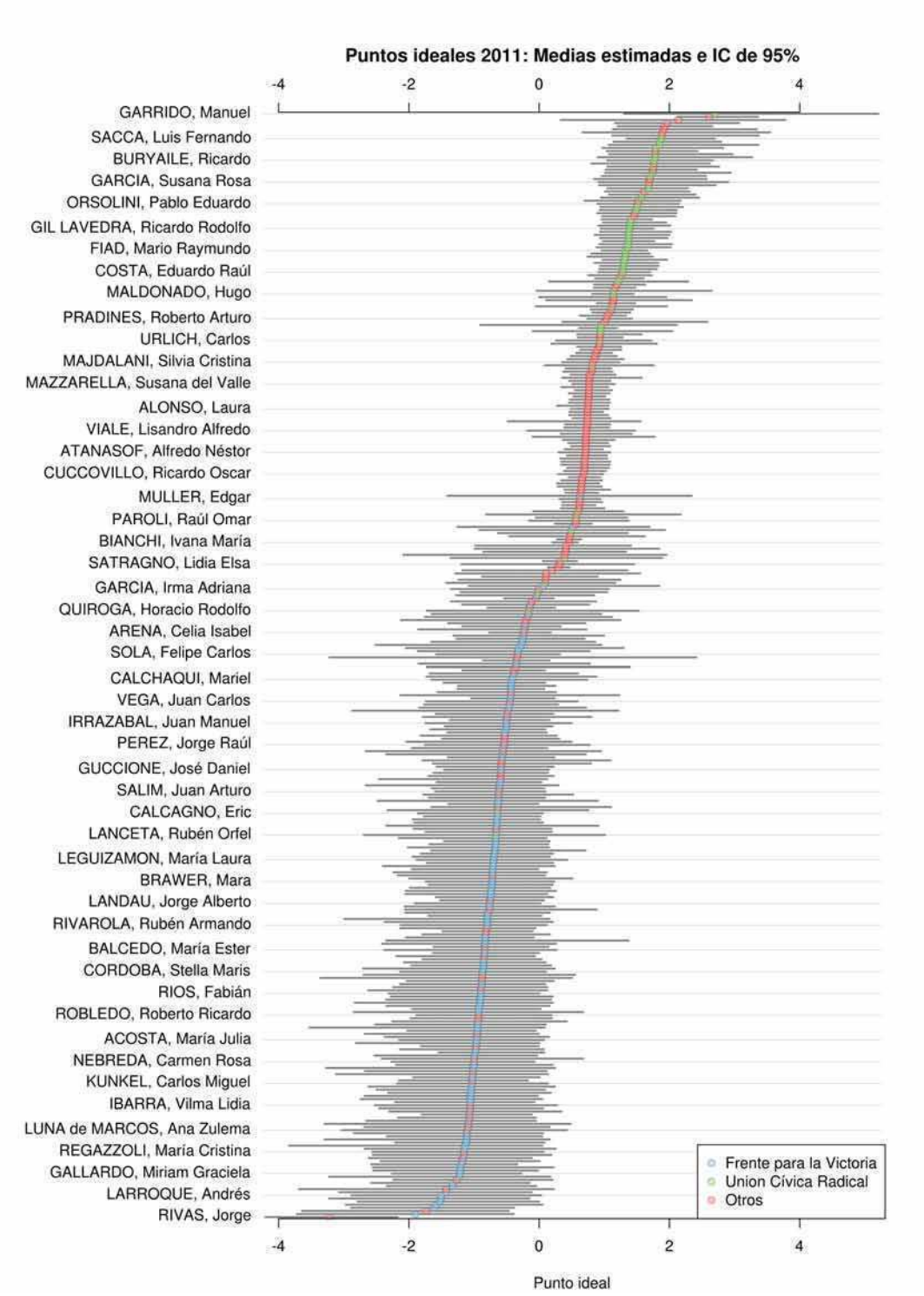

Figura 8: Puntos ideales del 2011, indicando intervalos de confianza del 95%

Efectivamente, como se ve en la Figura 8, los intervalos de confianza son en general mayores para todos los legisladores. Se deberá tener esto en consideración al interpretar resultados que involucren los puntos ideales de los años 2010 y 2011.

Cabe comparar la mayor cantidad de legisladores de punto ideal negativo de color rojo en la Figura 8 con la banda casi exclusivamente azul que conformaba el bloque oficialista en la Figura 7 correspondiente al 2006. Esto es consistente con el cambio en la conformación de la Cámara donde en el 2011 el oficialismo ya cuenta con partidos mayormente aliados tales como el Nuevo Encuentro o el Frente Cívico por Santiago, como se puede observar en la Figura 3.

Validación cruzada de las estimaciones de puntos ideales

Para verificar qué tan bien se puede predecir el comportamiento de los legisladores con las estimaciones que provee el algoritmo utilizado, se pueden realizar predicciones utilizando las probabilidades posteriores obtenidas, y contrastando con las votaciones efectivas de los legisladores.

Para esto se procede a realizar una validación cruzada de 10 cruces entre los datos, imputando como faltantes la décima parte de los votos nominales y obteniendo los estimadores con los datos restantes, para evitar entrenar el algoritmo con los mismos datos de testeo. Una vez predichas todas las votaciones, se calcula la tasa de acierto total para cada legislador. El código para esta validación cruzada se encuentra en `ideal_points/predict_ideal.R`.

Con estos datos se confeccionaron los gráficos de la Figura 9. Como se ve en la figura, los legisladores más difíciles de predecir son los de puntos ideales más cercanos al cero, por más que las Figuras 7 y 8 indican que los intervalos de confianza en esta región son en general más chicos. Los legisladores de posturas más extremas son más fáciles de predecir, por más que no se sepa bien si están en uno u otro lugar del oficialismo o la oposición. Más allá de eso, parece que en general el oficialismo es más fácil de predecir que la oposición, ya que los valores obtenidos para los legisladores con puntos ideales mayores a cero son en general más bajos que los de legisladores con puntos ideales menores a cero.

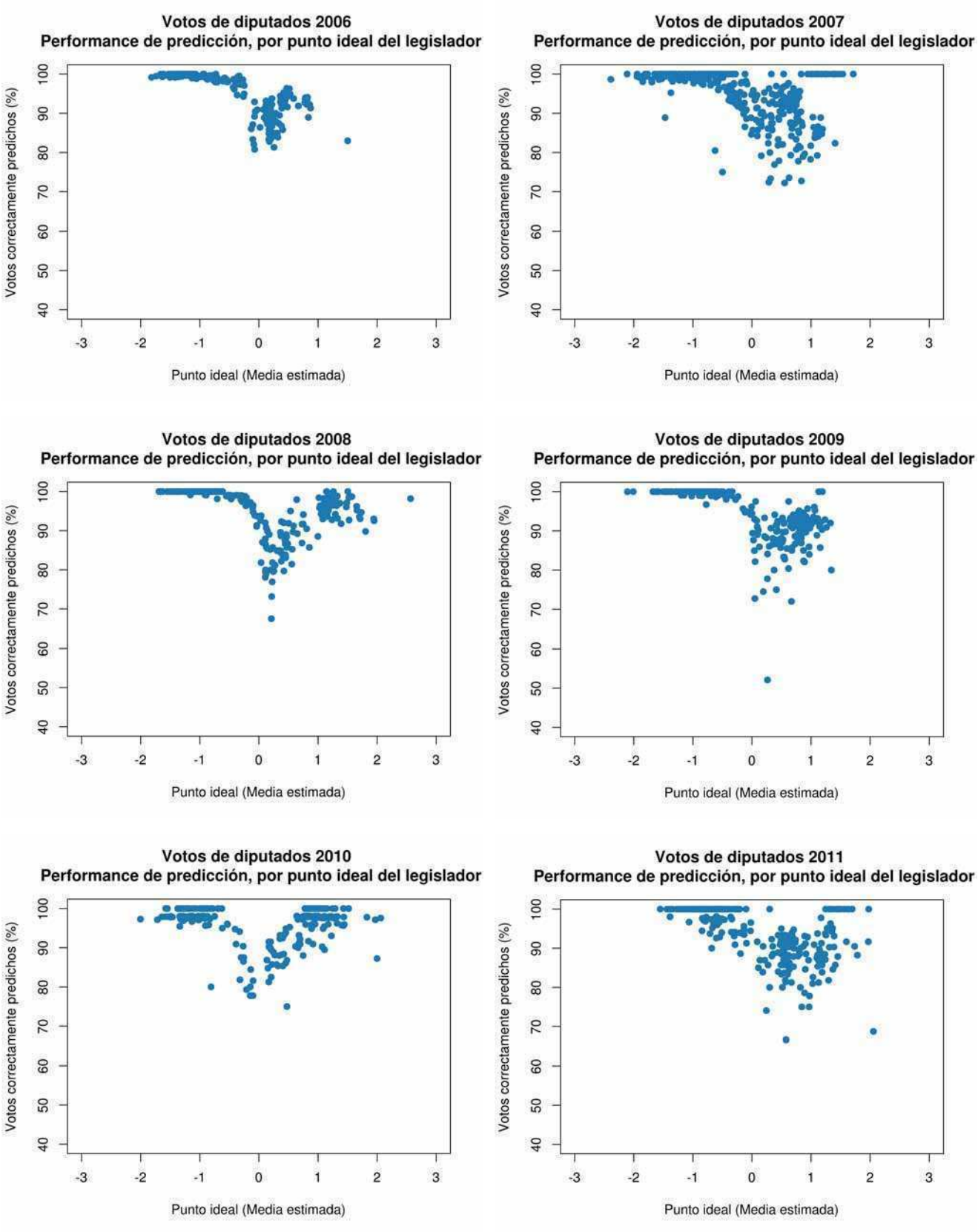

Figura 9: Performance de predicción utilizando los puntos ideales obtenidos

Tomado en general la efectividad de la predicción es bastante alta, si se utilizara como un clasificador. Se confeccionaron las matrices de confusión de predicción para los distintos años, como se puede ver en la Figura 10.

```
Datos de diputados 2006
        Predice No Predice Si
Vota No       5597        448
Vota Si       1379      21605
```

```
Datos de diputados 2007
        Predice No Predice Si
Vota No       3317        436
Vota Si        901      12833
```

```
Datos de diputados 2008
        Predice No Predice Si
Vota No       4196        205
Vota Si        696      10600
```

```
Datos de diputados 2009
        Predice No Predice Si
Vota No       2463        184
Vota Si        533      10008
```

```
Datos de diputados 2010
        Predice No Predice Si
Vota No       2715        115
Vota Si        266       4854
```

```
Datos de diputados 2011
        Predice No Predice Si
Vota No       1237         69
Vota Si        311       3648
```

Figura 10: Matrices de confusión de predicciones con los puntos ideales obtenidos

El desempeño del predictor es notablemente bueno, con una exactitud de entre 0.924 y 0.952 todos los años. Hay que tener en cuenta que si bien se realizó validación cruzada para no entrenar con los mismos datos con los que se testea, cada predicción se realiza “sabiendo” lo que votaron otros legisladores del mismo bloque o de puntos ideales cercanos durante la misma votación. El algoritmo no sería capaz de predecir el resultado de una votación antes de saber lo que votan al menos algunos legisladores, para poder estimar los parámetros del proyecto.

Normalización de los puntos ideales y evolución histórica

Un análisis interesante es observar cómo evolucionan los puntos ideales de los legisladores a lo largo de los distintos años con los que se trabaja.

Como se mencionó con anterioridad, antes de poder comparar puntos ideales de distintas corridas es necesario normalizar los resultados, debido a que cualquier inversión o escalado de los puntos ideales sería equivalente en términos de las votaciones observadas que generarían. Dado un conjunto de votaciones, dos corridas independiente del algoritmo podrían generar resultados en las que todos los puntos ideales cambian en un factor de escala constante, por ejemplo. Se debe imponer puntos de referencia suficientes para que los puntos ideales generados sean unívocos.

Para el caso en el que se trabaja con una sola dimensión, por suerte, es sencillo normalizar los puntos ideales para que tengan media cero y desviación estándar uno, con lo cual se elimina toda variación posible, salvo la inversión: dado un juego de puntos ideales con media cero y desvío estándar uno que satisface un conjunto de votaciones, si se cambia el signo de todos los puntos ideales se obtiene otro juego de puntos ideales que sigue maximizando la utilidad de los legisladores para ese conjunto de votaciones. Para el presente trabajo, si se agrega la suposición de que los puntos ideales oficialistas son los de valor menor que cero, se pueden utilizar los resultados para realizar comparaciones entre años.

La Figura 11 muestra los distintos valores de punto ideal normalizado para los 39 legisladores que tuvieron una banca durante los 6 años analizados.

	2006	2007	2008	2009	2010	2011	Plot	Media	Max Diff
ALVAREZ, Juan José	0.7	0.52	1.7	1.1	-0.13	-0.86		0.51	2.56
BERTONE, Rosana Andrea	-1.09	-0.85	-1.01	-1.08	-0.9	-1.09		-1	0.24
BRUE, Daniel Agustín	0.43	0.35	-0.14	-0.66	-1.39	-0.71		-0.35	1.82
CAMAÑO, Graciela	0.21	-1.38	0.03	0.19	0.87	0.77		0.12	2.25
CIGOGNA, Luis Francisco J.	-1.15	-1.33	-1.06	-0.83	-1.33	-0.73		-1.07	0.6
COMELLI, Alicia Marcela	0.8	0.72	0.17	0.13	-0.22	-1.03		0.09	1.83
CONTI, Diana Beatriz	-1.01	-0.79	-0.4	-0.8	-0.94	-0.6		-0.76	0.61
CORDOBA, Stella Maris	-0.45	-2	-0.28	-1.21	-1.25	-0.78		-1	1.72
DAHER, Zulema Beatríz	-0.66	-0.12	0.14	0.66	0.77	0.37		0.19	1.43
DE NARVAEZ, Francisco	0.68	0.56	1.27	1.1	1.19	0.68		0.91	0.71
DEPETRI, Edgardo Fernando	-1.17	-0.97	-0.15	-1.48	-0.69	-0.62		-0.85	1.34
DIAZ BANCALARI, José M.	0.15	-2.13	-0.68	-0.77	-1.35	-0.92		-0.95	2.28
DIAZ ROIG, Juan Carlos	-0.53	-0.1	-1.19	-0.93	-1.19	-0.91		-0.81	1.1
DIAZ, Susana Eladia	-0.51	-0.57	-1.15	-1.08	-0.97	-0.8		-0.84	0.64
di TULLIO, Juliana	-0.96	-1.15	-0.79	-1.37	-1.37	-1.26		-1.15	0.58
FADEL, Patricia Susana	-1.32	-1.66	-0.77	-1.02	-1.04	-0.81		-1.1	0.9
FIOL, Paulina Esther	-0.91	-1.26	-0.93	-1.68	-1.58	-0.78		-1.19	0.89
GARCIA, Susana Rosa	1.63	1.43	1.5	1.43	1.04	1.64		1.44	0.61
GIOJA, Juan Carlos	-1.02	-0.88	-0.86	-0.72	-1.44	-0.82		-0.96	0.72
GIUBERGIA, Miguel Angel	1.07	0.89	1.33	1.22	1.03	1.31		1.14	0.44
GODOY, Ruperto Eduardo	-1.41	-1.19	-0.97	-1.05	-1.53	-0.73		-1.15	0.8
GONZALEZ, Nancy Susana	-1	-1.43	-1.75	-0.54	-1.32	-0.51		-1.09	1.23
IRRAZABAL, Juan Manuel	-0.43	-0.05	-0.86	-0.75	-1.1	0.06		-0.52	1.16
KUNKEL, Carlos Miguel	-1.11	-1.35	-1.14	-1.34	-1.14	-0.53		-1.1	0.83
LANDAU, Jorge Alberto	-0.18	-1.72	-0.86	-1.03	-1.03	-1.11		-0.99	1.54
LOZANO, Claudio Raúl	1.03	1.12	0.5	0.74	0.21	0.67		0.71	0.91
MACALUSE, Eduardo Gabriel	1.62	1.42	0.4	0.77	0.19	1.92		1.05	1.73
MARCONATO, Gustavo A.	-1.41	-1.34	-1.04	-0.75	-2.41	-0.13		-1.18	2.28
MORENO, Carlos Julio	-1.64	-0.84	-1.09	-0.92	-1.41	-1.09		-1.17	0.79
PERIE, Hugo Rubén	-0.52	-0.54	-0.8	-1.12	-1.09	-0.48		-0.76	0.64
PINEDO, Federico	0.77	0.99	1.13	1.04	0.73	0.51		0.86	0.62
QUIROZ, Elsa Siria	1.59	1.43	1.55	1.49	1	1.63		1.45	0.63
RECALDE, Héctor Pedro	-1.16	-1.44	-1.22	-0.82	-1.39	-1.03		-1.18	0.62
RODRIGUEZ, Marcela V.	1.7	1.46	1.56	1.64	0.46	1.5		1.39	1.25
ROSSI, Agustín Oscar	-0.97	-1.12	-0.81	-0.79	-1.26	-1.05		-1	0.47
SALIM, Juan Arturo	-0.83	-0.65	-1.16	-0.94	-1.17	-0.78		-0.92	0.52
SLUGA, Juan Carlos	-1.06	-1.14	-0.89	-1.27	-0.83	-0.25		-0.91	1.02
THOMAS, Enrique Luis	-0.85	-0.22	0.24	0.87	1.31	0.67		0.34	2.16
WEST, Mariano Federíco	-1.09	-1.34	-1.29	-0.69	-1.23	-1.39		-1.17	0.7

Figura 11: Evolución histórica de los puntos ideales de los legisladores

El código para generar estos gráficos se encuentra en el archivo `historic_evolution/evolution.R`.

Es interesante mirar en esta figura los casos como Graciela Camaño, Enrique Thomas o Juan José

Álvarez, que no sólo tuvieron una gran variación en su postura legislativa sino que tuvieron un cambio de signo. Si se lee la historia de estos legisladores se descubre que han cambiado de bloque o de relación con el gobierno a lo largo de los años extraídos.

Evolución de la polarización de las posturas legislativas

Un resultado más que se puede obtener con los puntos ideales normalizados es la evolución de la polarización de los mismos, es decir cuantificar el progreso de la brecha entre oficialismo y oposición. Para esto se utiliza nuevamente la librería `mixtools` de R para ajustar un modelo de mezcla de 2 normales a los datos, y se muestra en la Figura 12 la evolución de la distancia entre las medias de las dos curvas.

El código utilizado se encuentra en el archivo `ideal_points/polarizacion.R`.

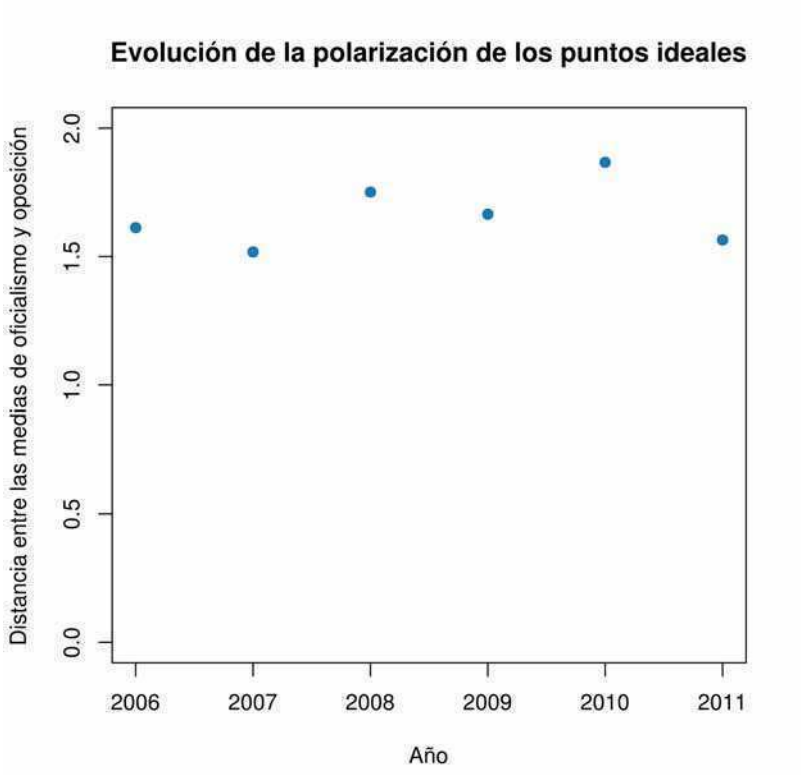

Figura 12: Evolución de la polarización de posturas legislativas

Si bien la curva tiene un máximo en el año 2010, no hay un cambio significativo en los años

adquiridos, ni parece haber una tendencia clara en una dirección o en otra.

Comparación de posturas legislativas: Recalde, Macaluse y Quiroz

Antes de comenzar a agrupar legisladores de posturas legislativas similares en la siguiente sección, vale la pena mirar algún caso puntual más en detalle para apreciar el tipo de fenómenos que ocurren según la similitud de posturas legislativas entre legisladores.

Para esto se presenta un estudio un poco más detallado de las votaciones de tres legisladores que estuvieron en la cámara durante todos los años analizados, y mantuvieron un punto ideal relativamente constante en su postura:

- Héctor Recalde (IP medio -1.18, máxima variación 0.62), peronista desde su juventud, diputado por Frente para la Victoria durante los seis años analizados.
- Eduardo Macaluse ( IP medio 1.05, máxima variación 1.73), en el 2006 ocupaba una banca por el ARI y en el 2007 es reelecto por el mismo partido, pero luego forma su propio bloque Solidaridad e Igualdad.
- Elsa Quiroz (IP medio 1.43, máxima variación 0.63), legisladora por el ARI, ahora Coalición Cívica – ARI, y secretaria del mismo partido.

De esta manera tenemos un legislador oficialista y dos de oposición. Cuando se compara cómo votaron estos legisladores se ve que efectivamente Macaluse y Quiroz coinciden en las votaciones notablemente más que con Recalde, como se ve en la Figura 13. No obstante, incluso los legisladores de posturas opuestas votan igual la mayoría de las veces; no es que dos legisladores de posturas claramente opuestas se contradigan siempre en las votaciones.

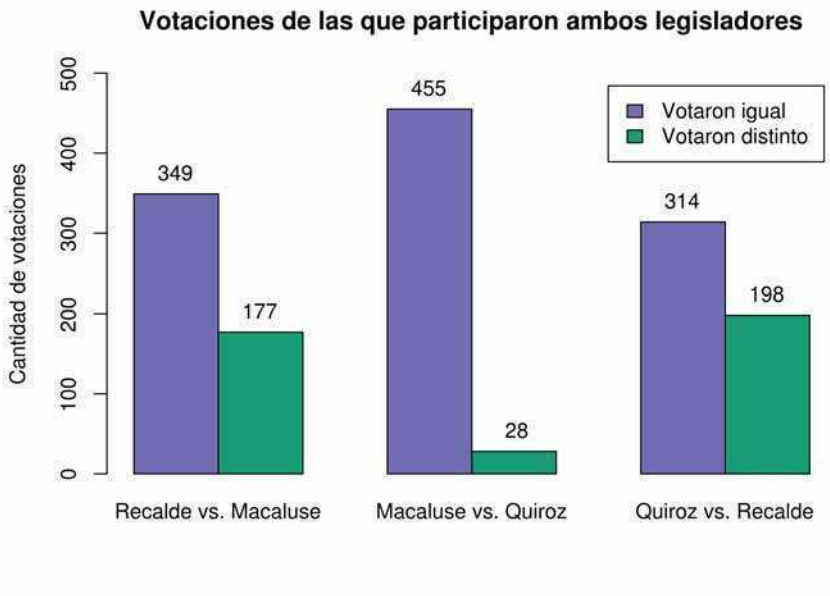

Figura 13: Cantidad de veces que coincidieron, o no, los distintos pares de legisladores

Para visualizar los tipos de proyectos en los que estos legisladores votaron de una u otra manera, se extrajeron los textos y fundamentos de aquellos proyectos en los que estos legisladores votaron, y los términos más relevantes se presentan en la Figura 14.

En la Figura 14 a la izquierda se muestran los términos más relevantes de los proyectos en los que los tres legisladores coincidieron en la votación. A la derecha, los términos más relevantes de los proyectos en los que Macaluse y Quiroz coincidieron en votar de una forma, y Recalde votó de otra.

Los textos fueron lematizados, y el peso de cada término es el tf-idf dentro del corpus de los proyectos visualizados. Por más detalles sobre la lematización ver la sección *Preprocesamiento y lematización*.

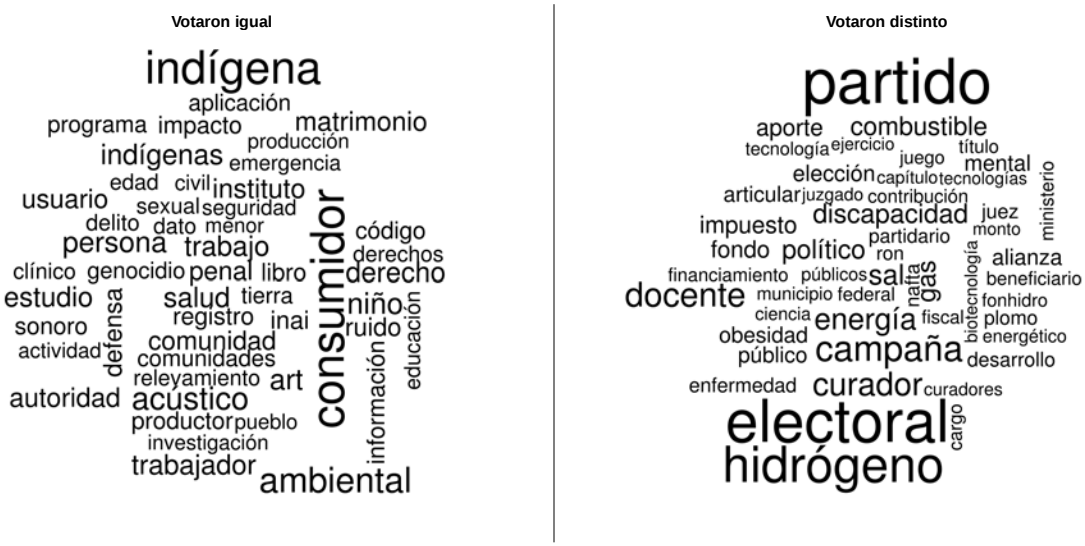

Figura 14: Términos más relevantes en proyectos en los que los tres legisladores coincidieron, o no.

Parecería que en temas sociales y de derechos humanos los legisladores tienden a coincidir, mientras que en temas económicos o políticos hay mayor debate, pero esto es una interpretación completamente personal. No se realizaron análisis cuantitativos al respecto.

El código escrito para esta sección se encuentra en la carpeta `comparacion/`.

Agrupamiento y segmentación de legisladores

Ya con conocimiento de qué tipo de similitudes existen entre los comportamientos de distintos legisladores, se proponen tres técnicas de agrupamiento para determinar si los bloques políticos determinan el comportamiento, o si se pueden detectar alianzas entre legisladores o quiebres dentro de un bloque:

- Búsqueda de cliques en el grafo determinado por la relación “concuerda con”.
- Agrupamiento jerárquico.
- Clustering mediante el algoritmo de k-medias.

Para darle un contexto un poco más acotado a este análisis, en esta sección se trabaja únicamente con los datos del 2009.

Búsqueda de cliques en el grafo de la relación “concuerda con”

Como el conjunto de datos es suficientemente pequeño, se intenta una aproximación de búsqueda exhaustiva para encontrar grupos de legisladores de comportamiento similar.

En base al análisis de las votaciones de legisladores del punto anterior, en principio parecería que dos legisladores de posturas opuestas no van a votar siempre en contra, pero dos legisladores extremadamente alineados pueden llegar a votar siempre igual, o a lo sumo a no contradecirse. Se define entonces la relación “concuerda con” de la siguiente manera:

> Un legislador “concuerda” con otro, si en toda votación en la cual ambos efectivamente votaron por Afirmativo o Negativo, votaron igual.

Es fácil ver que esta relación es reflexiva y simétrica, pero no transitiva.

Se define entonces un grafo no dirigido con un vértice por legislador, y se establece una arista entre dos vértices siempre que los legisladores concuerden. Se adjunta una representación del grafo en formato Graphviz[8] en el archivo `clusters/full.dot`.

Este grafo es en principio bastante denso, con 28.6 aristas por vértice en promedio, por lo que antes de emprender una búsqueda exhaustiva se filtran algunos nodos manualmente.

Analizando los legisladores que votaron por Negativo una sola vez, aparecen los legisladores que votaron por la negativa sólo para la solicitud de tratamiento sobre tablas del expediente 5535-D-09, un proyecto de la oposición. Sumando a esto los legisladores que siempre que votaron lo hicieron por Afirmativo, pero se encontraban ausentes para la votación de esta solicitud de tratamiento sobre tablas, resulta en un clique de 105 legisladores de punto ideal entre -1.84 y -0.28, por lo que conforma la gran mayoría del oficialismo. No se incluye aquí el listado de los integrantes por cuestiones de espacio, pero se adjunta en el archivo `clusters/cliques.txt`.

Una vez quitados estos legisladores, se procede a realizar una búsqueda exhaustiva sobre los legisladores restantes, con los siguientes resultados, en orden decreciente de tamaño del clique:

8 http://www.graphviz.org/

```
* Aguad - Unión Cívica Radical (0.81)
* Azcoiti - Unión Cívica Radical (0.84)
* Barrionuevo - Justicialista (0.93)
* Bayonzo - Unión Cívica Radical (1.12)
* Cusinato - Unión Cívica Radical (0.81)
* Fabris - Frente de Todos (0.66)
* Giubergia - Unión Cívica Radical (0.92)
* Guerci De Siufi - Unión Cívica Radical (1.11)
* Kroneberger - Unión Cívica Radical (0.66)
* Lanceta - Unión Cívica Radical (0.85)
* Lemos - Unión Cívica Radical (1.02)
* Martinez Garbino - Nuevo Espacio Entrerriano (0.69)
* Martinez Oddone - Unión Cívica Radical (1.02)
* Morandini - Alianza Frente Nuevo (1.14)
* Nieva - Unión Cívica Radical (0.66)
* Poggi - Justicialista (0.91)
* Portela - Unión Cívica Radical (0.72)
* Rioboo - Unión Cívica Radical (0.83)
* Sola - Unión Pro (0.41)
* Storni - Unión Cívica Radical (1.11)
* Urlich - Unión Cívica Radical (0.75)
* Del Campillo - UNA (0.76)

* Chiquichano - Justicialista (-0.77)
* Garcia De Moreno - Frente para la Victoria (-0.87)
* Gonzalez - Justicialista (-0.58)
* Gonzalez - Frente para la Victoria (-0.57)
* Pais - Frente para la Victoria (-0.67)

* Cremer De Busti - Frente Justicialista para la Victoria (0.44)
* Petit - Frente para la Victoria (0.55)
* Zavallo - Frente Justicialista para la Victoria (0.42)

* Cigogna - Justicialista (-1.05)
* Giannettasio - Frente para la Victoria (-1.09)
* Morante - Alianza Frente Justicialista Chaco Merece Mas (-0.70)
```

Finalmente, se encuentran 40 cliques de tamaño 2, compuestos en su gran mayoría por legisladores de oposición. En todos los casos es claro cómo los cliques están compuestos por legisladores de punto ideal similar, como era de esperarse. Pero además, resulta notable que el oficialismo forma cliques de gran tamaño, cubriendo la mayoría de los diputados con un solo clique, mientras que la oposición se encuentra fragmentada en múltiples posturas legislativas sutilmente diferentes. El listado completo de los cliques maximales encontrados se adjunta en el archivo `clusters/cliques.txt`.

Para visualizar este fenómeno se realiza la Figura 15, que muestra el tamaño de los cliques según el punto ideal promedio dentro del grupo. El eje vertical indica el desvío estándar del punto ideal dentro del grupo, que da una idea de la variabilidad de los legisladores que integran el grupo.

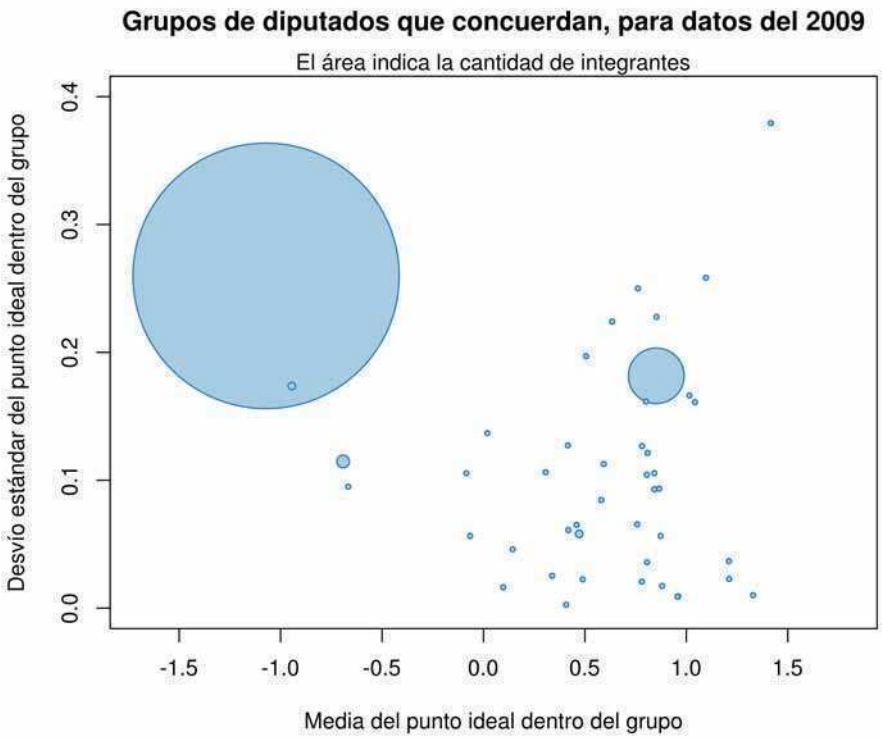

Figura 15: Tamaño de los grupos de diputados que concuerdan, según el punto ideal.

En esta figura se aprecia tanto la baja variabilidad del punto ideal dentro de los grupos, incluso para los grupos más grandes, como así también la mayor fragmentación de los diputados de puntos ideales mayores a cero.

Agrupamiento Jerárquico

Para aplicar las técnicas más conocidas de agrupamiento entre legisladores, tanto para el algoritmo de agrupamiento jerárquico como para k-medias es necesario definir una medida de distancia entre legisladores.

Para este fin, se sitúan a los legisladores en un espacio N-dimensional donde N es la cantidad de votaciones realizadas en el año. A cada legislador, por cada dimensión, se le asigna el valor 1 si para dicha votación el diputado votó por Afirmativo, -1 si votó por Negativo, y 0 si estuvo ausente o votó por Abstención.

Se consideran dos alternativas a imputar como 0 los Ausentes y Abstenciones. Por un lado se podría

acercar un poco el valor de Ausente al Negativo, en vistas de los resultados de la Tabla 2 en la que se ve que legisladores frecuentemente utilizan el Ausente como un recurso para evitar que un proyecto se apruebe, por lo que se podría asignar un valor pequeño pero negativo al Ausente. Dado que es complejo analizar cuándo un legislador está ausente porque no quiere que se apruebe un proyecto y cuando está ausente por otras razones, se descartó esta alternativa. Por otro lado, se podrían imputar los ausentes como datos faltantes utilizando los valores predichos según el punto ideal, tomados de la validación de los puntos ideales (ver sección *Validación de las estimaciones de puntos ideales*). Esta alternativa podría ser interesante, pero no se implementa por cuestiones de tiempo y prioridades.

Estableciendo entonces esta medida de distancia, antes de realizar el agrupamiento jerárquico se puede visualizar la distancia de cada uno de los legisladores a los otros legisladores, como se muestra en la Figura 16.

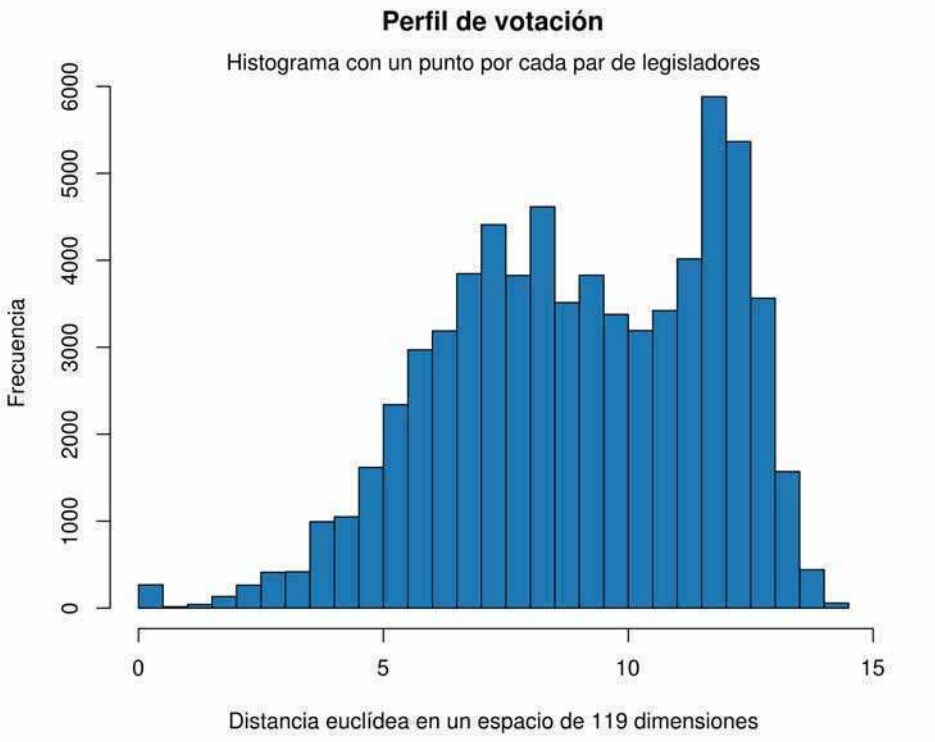

Figura 16: Distancia entre cada par de legisladores del 2009

Es notable que con esta definición, las distancias entre legisladores adoptan el mismo perfil bi-modal

que se mostró en la Figura 2, de manera consistente. Aquí se puede interpretar como que si se toma un par de legisladores de cualquier postura legislativa, es más probable que sea un par de “aliados” o “adversarios”, que legisladores de una distancia intermedia.

Una vez ubicados los legisladores en este espacio, para el agrupamiento jerárquico se utiliza sencillamente la distancia euclídea como medida de distancia, y se prueban distintos criterios de agrupamiento. En la Figura 17 se muestran dos fragmentos de los resultados obtenidos utilizando el método de Ward [Everitt 2011], el cual en cada paso une los dos grupos que menos incrementen la varianza total intra-grupo. Nótese que esto es considerando las distancias euclídeas definidas, sin mirar el punto ideal del legislador en absoluto. Esto resultó en grupos compactos, con puntos ideales más similares y mayor proporción de legisladores de un mismo partido, que utilizando los otros criterios de enlace simple o enlace completo. No se muestra el dendrograma completo por cuestiones de espacio, pero el mismo está disponible en los archivos adjuntos en `hclust/hclust_2009.png`.

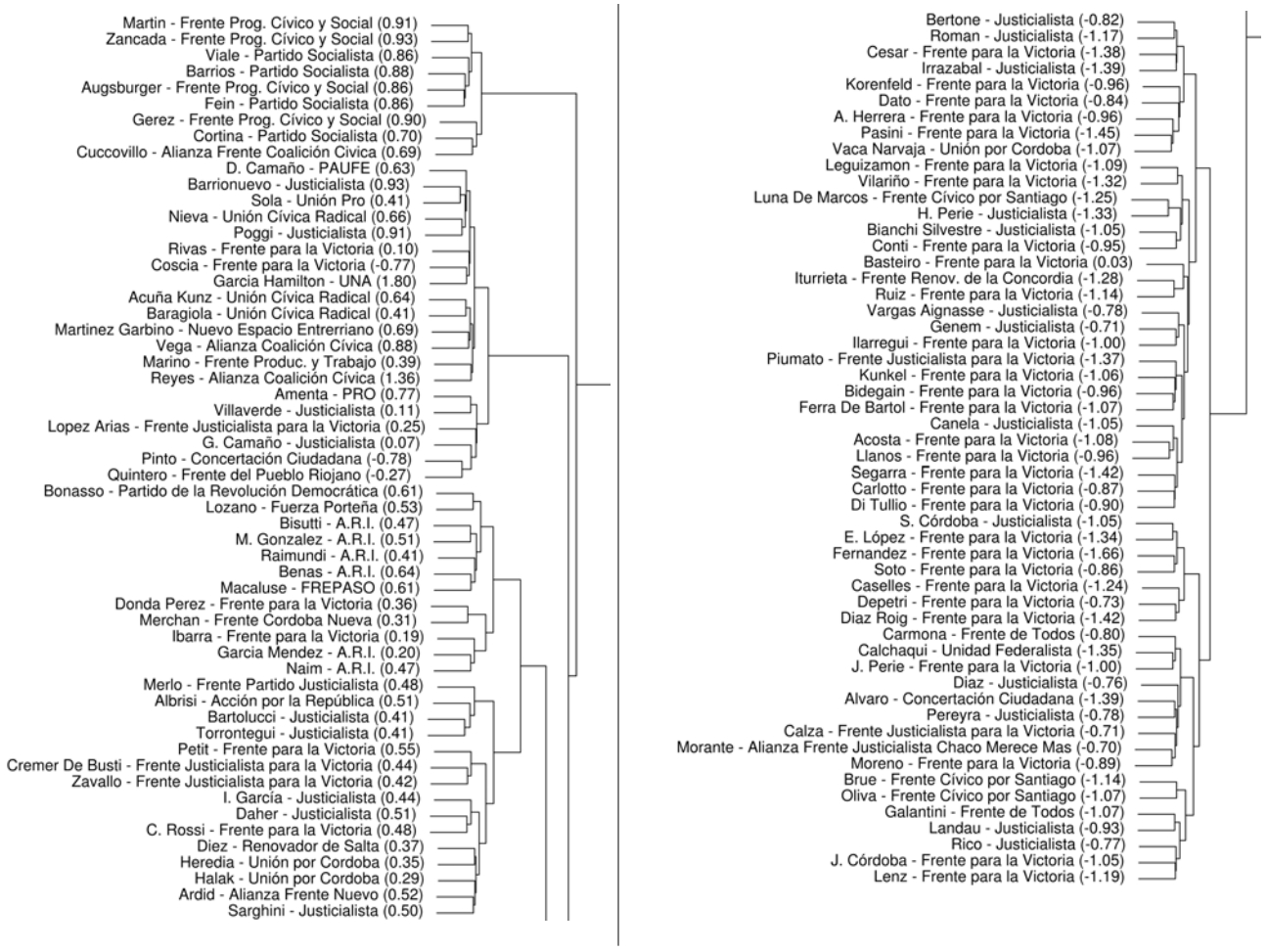

Figura 17: Dos fragmentos del dendrograma mediante el método de Ward, para datos del 2009.

Si bien el agrupamiento jerárquico separa efectivamente en un primer nivel lo que es oficialismo de oposición, cuando se observa la conformación de los grupos más pequeños se encuentra una mayor variedad en cuanto a los bloques políticos de los legisladores que los conforman, salvo algunos grupos menores claramente definidos. Esta técnica tiene además el problema adicional de que el corte de los grupos queda sin resolver, por lo que es necesario algún criterio para definir exactamente cuáles son los grupos resultantes.

Para comparar con el método anterior se generó el mismo gráfico de la Figura 15, con los tamaños de cada grupo, el punto ideal promedio y la variabilidad intra-grupo, tomando cortes del dendrograma en 6 y 12 grupos, y se muestra en la Figura 18.

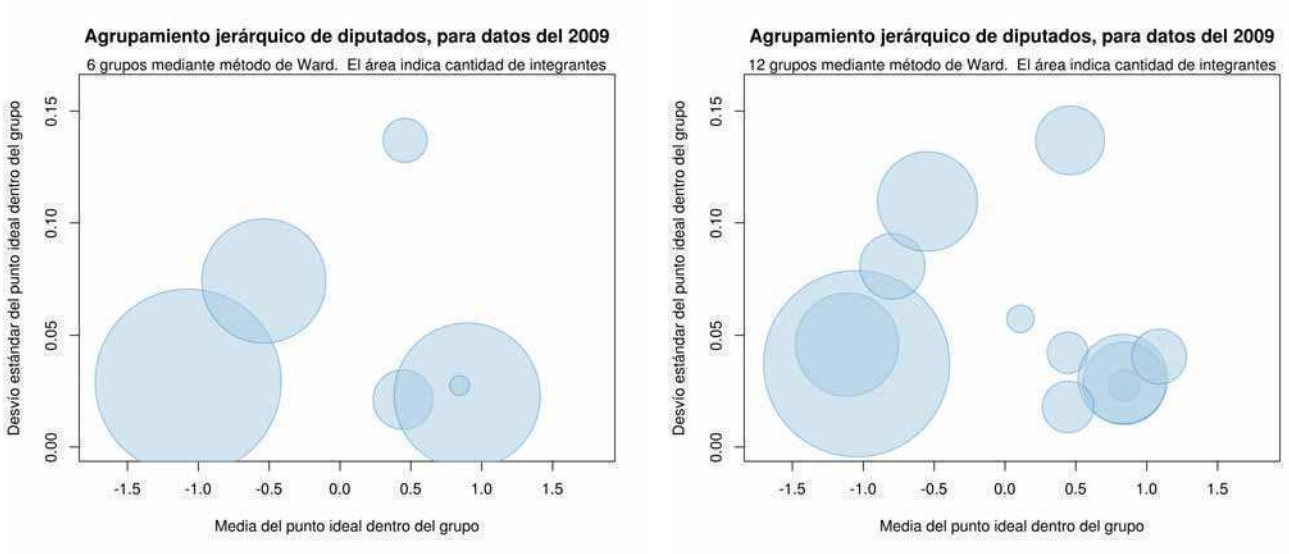

Figura 18: Tamaño del grupo según el punto ideal promedio de los integrantes

Si bien el método marca menos la diferencia entre grupos grandes y pequeños, es notablemente buena la baja variabilidad interna de los puntos ideales de los legisladores en cada grupo. Por un lado el método de Ward optimiza precisamente la variabilidad interna de cada grupo, pero eso es basado en las distancias en el espacio euclídeo definido para el agrupamiento, sin mirar los puntos ideales de los legisladores.

Agrupamiento mediante k-medias

Finalmente, se utilizó el algoritmo de k-medias para segmentar los legisladores, utilizando la misma medida de distancia que para el algoritmo de agrupamiento jerárquico.

Se realizaron varias corridas para distinto número de grupos, para obtener entre tres y doce grupos en total. En las Figuras 19 y 20 se muestra la conformación de los grupos para 6 y 12 grupos, según cantidad de integrantes de cada bloque. Todos los resultados se encuentran en los archivos adjuntos, en la carpeta kmeans/.

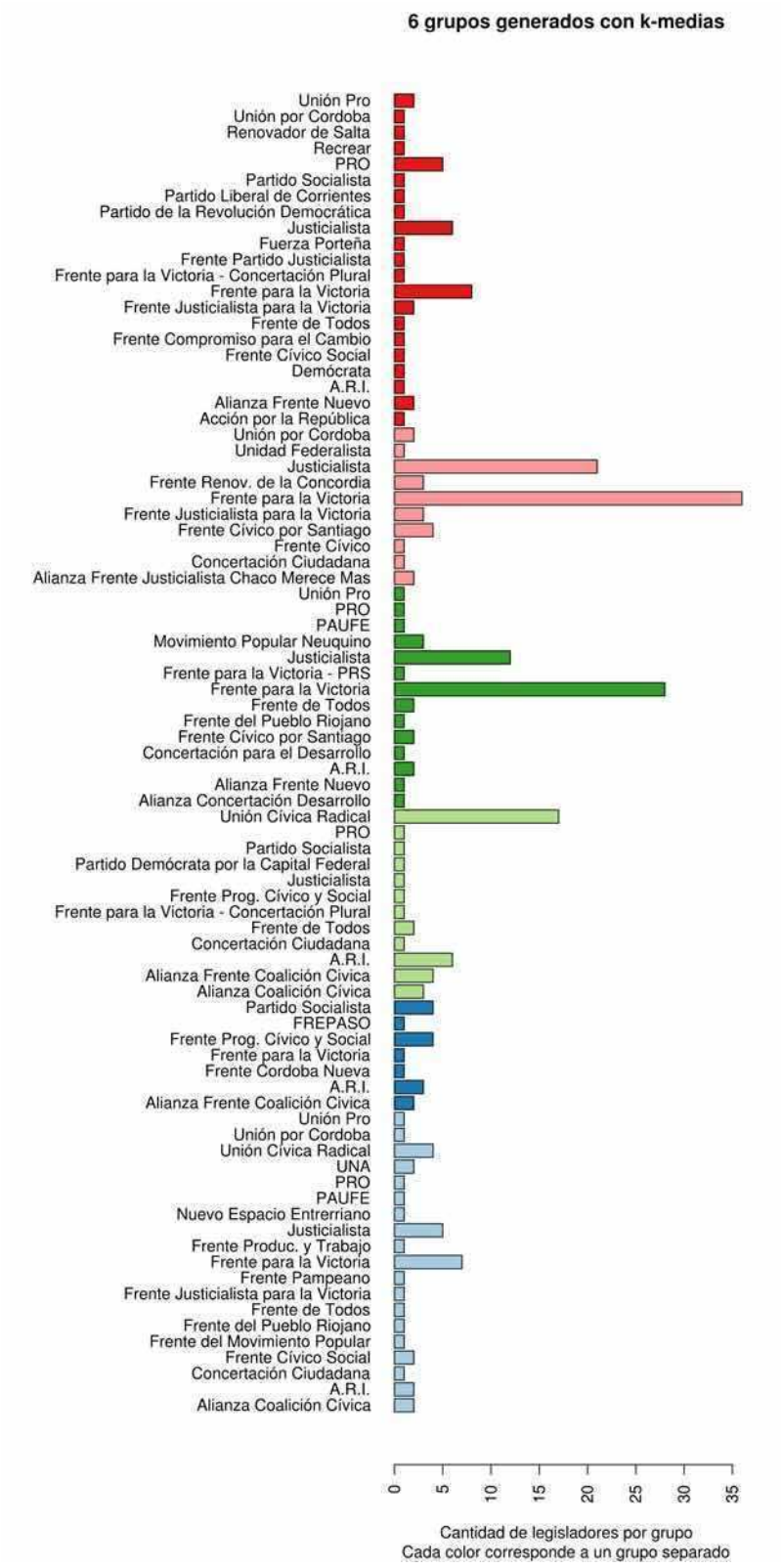

Figura 19: Conformación de 6 grupos generados con K-medias, con datos del 2009

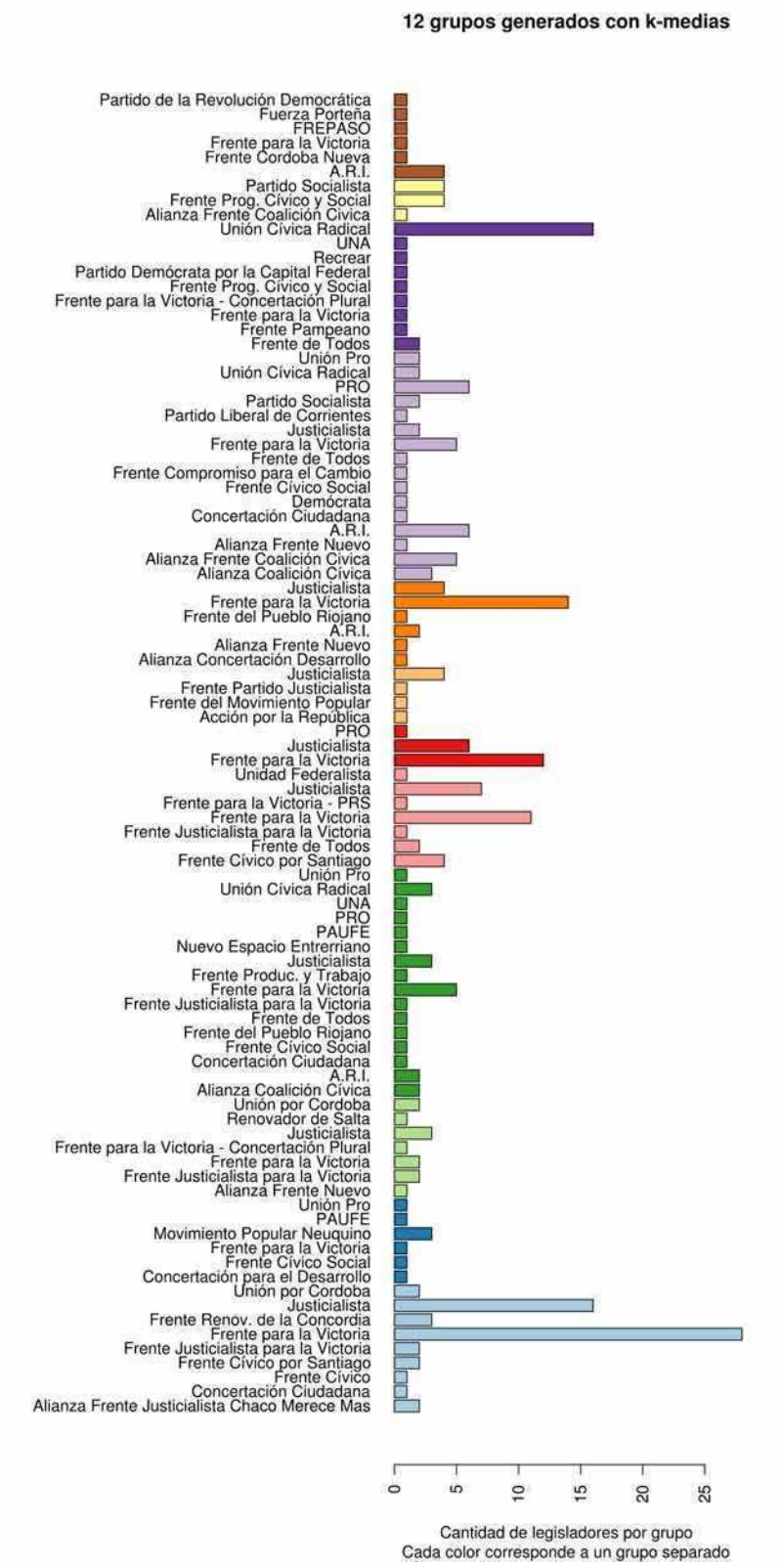

Figura 20: Conformación de 12 grupos generados con K-medias, con datos del 2009

Los grupos generados en ambos casos son de un tamaño más homogéneo que los cliques de legisladores que concuerdan, pero también tienen más cantidad de legisladores de distintos partidos políticos, por lo que se supondría que son de una varianza interna más alta. Incluso en el caso de 12 grupos, todos los grupos están conformados de al menos tres partidos distintos. Por otro lado si se quisiera obtener grupos de tamaño homogéneo seguramente sería necesario empezar a unir partidos de esta manera. En el caso de K-Medias además interviene la selección inicial de los centros de grupo, que para el código utilizado es al azar.

Para comparar mejor los resultados aquí obtenidos con los presentados en las secciones anteriores se confeccionaron los mismos gráficos de burbujas que muestran el tamaño de cada grupo según el punto ideal promedio y el desvío estándar dentro del grupo, que se muestra en la Figura 21.

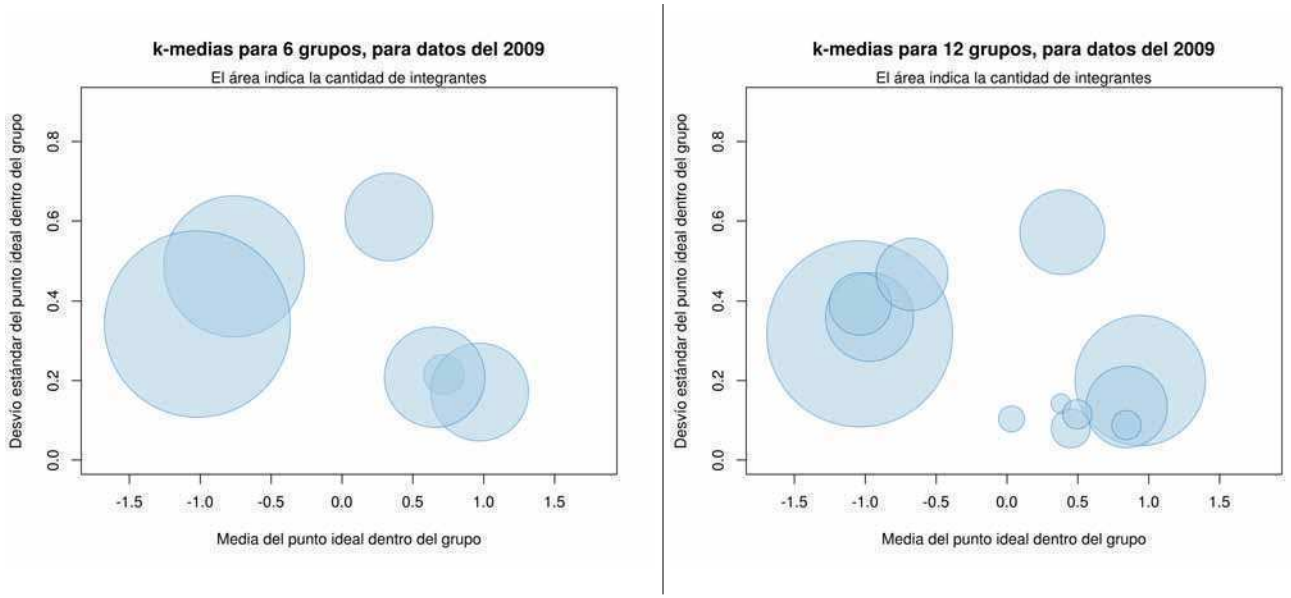

Figura 21: Tamaño del grupo según punto ideal promedio de los integrantes

Efectivamente se ve que la variabilidad interna de cada grupo es mayor utilizando k-medias, y el tamaño de los grupos es más homogéneo.

Si hubiera que elegir un algoritmo para agrupar legisladores habría que ver el propósito del agrupamiento. Si lo que se necesita es detectar las verdaderas alianzas y acuerdos entre legisladores, probablemente el algoritmo de búsqueda de cliques sea el más adecuado. Si lo que se necesita es mirar tendencias generales, o convergencias entre partidos, o segmentar a los legisladores en cierta cantidad

fija de grupos relativamente homogéneos, el algoritmo de agrupamiento jerárquico por el método de Ward parece ser el más indicado.

Fidelidad al partido mediante análisis de la entropía

Un estudio interesante que hace [Sirovich 2003] es el análisis de las votaciones de la Corte Suprema de Justicia de los Estados Unidos del juez Rehnquist entre los años 1994 y 2003, desde el punto de vista de la teoría de la información.

La Corte Suprema de Justicia de los Estados Unidos está compuesta por 9 jueces, por lo que considera una votación como un vector de nueve dígitos ±1, y sin pérdida de generalidad establece que +1 es el voto de la mayoría, por lo que hay sólo 2^8 votaciones posibles. En términos geométricos, el conjunto de votaciones se ubica en un espacio euclídeo 9-dimensional restringido al semi-espacio positivo sobre la "esfera de decisiones" de radio unitario,

$$\sum_j n_j^2 = 9, j \in [1..9]$$

Aquí los n_j son los votos de cada uno de los 9 jueces de la Corte, $n_j = \pm 1$. Como se asume que +1 es el voto mayoritario, se tiene

$$\sum_j n_j > 0$$

Sirovich plantea dos casos extremos como simplificaciones de la Corte que sirven como ejemplo:

- La corte "platónica", donde los nueve jueces votan al azar, y en cada decisión todas las votaciones son igual de probables. Por ende en cada votación hay 8 bits de información.
- La corte "omnisciente", donde todos los jueces son tan sabios que siempre están de acuerdo, y las votaciones son siempre unánimes, iguales al vector [1,1,...,1]. En cada decisión hay 0 bits de información, y la corte se comporta como si estuviera compuesta de un sólo juez platónico.

Así explica que la cantidad de "jueces platónicos equivalentes" en una corte es la cantidad de bits de información promedio en cada votación, más uno. Aquí la cantidad de bits de información (o entropía) en cada votación está dada por la definición de la información de Shannon,

$$I = -\sum_b p_n \log_2(p_n)$$

Si se realiza este mismo análisis dentro de cada bloque de legisladores es posible medir el efecto de la fidelidad al partido. Se puede suponer que el partido en sí mismo plantea un punto ideal legislativo, y cada diputado (que de ser independiente tendría una postura distinta) se acerca en mayor o menor medida al punto ideal de su bloque.

Hay dos suposiciones necesarias para aplicar este análisis a las votaciones de diputados. En primer lugar, en caso de que haya un empate entre las votaciones de los legisladores de un bloque, se considera que gana el voto por Afirmativo. A diferencia de la Corte Suprema que siempre consta de un número impar de jueces para evitar los empates, los bloques de diputados pueden ser de una cantidad par de legisladores, y puede darse un empate en caso de desacuerdo. Esto no afecta los resultados y es simplemente para poder establecer unívocamente el valor de los vectores de votación. Por otro lado, los Ausentes y Abstenciones se consideran votos a favor de la mayoría dentro del bloque. Esta modificación hace que en los resultados los legisladores aparenten votar más en bloque de lo que en realidad hacen. Finalmente, para cada año sólo se analizan los bloques compuestos por más de un legislador.

Además hay otro efecto que interviene, que es la distancia original entre el punto ideal del partido y el del legislador, pero es razonable suponer que cada partido es aproximadamente igual de exitoso postulando legisladores lo más afines posible al punto ideal del bloque en sí, por lo que se asume que una menor entropía es debida a la fidelidad al partido, por más que se sepa que en realidad se miden ambos efectos a la vez.

Como los bloques más grandes naturalmente van a contener una mayor entropía en cada votación debido al mayor número de legisladores, para poder comparar un bloque con otros es necesario normalizar los resultados respecto del tamaño del bloque. Para ello se divide la información en bits obtenida para cada bloque por el tamaño del bloque, para obtener la información contenida en cada voto de cada legislador del bloque.

Los resultados se adjuntan en los archivos `shannon/entropia_intra_bloque_*.csv`, y se muestran en la Figura 22, por cada año analizado.

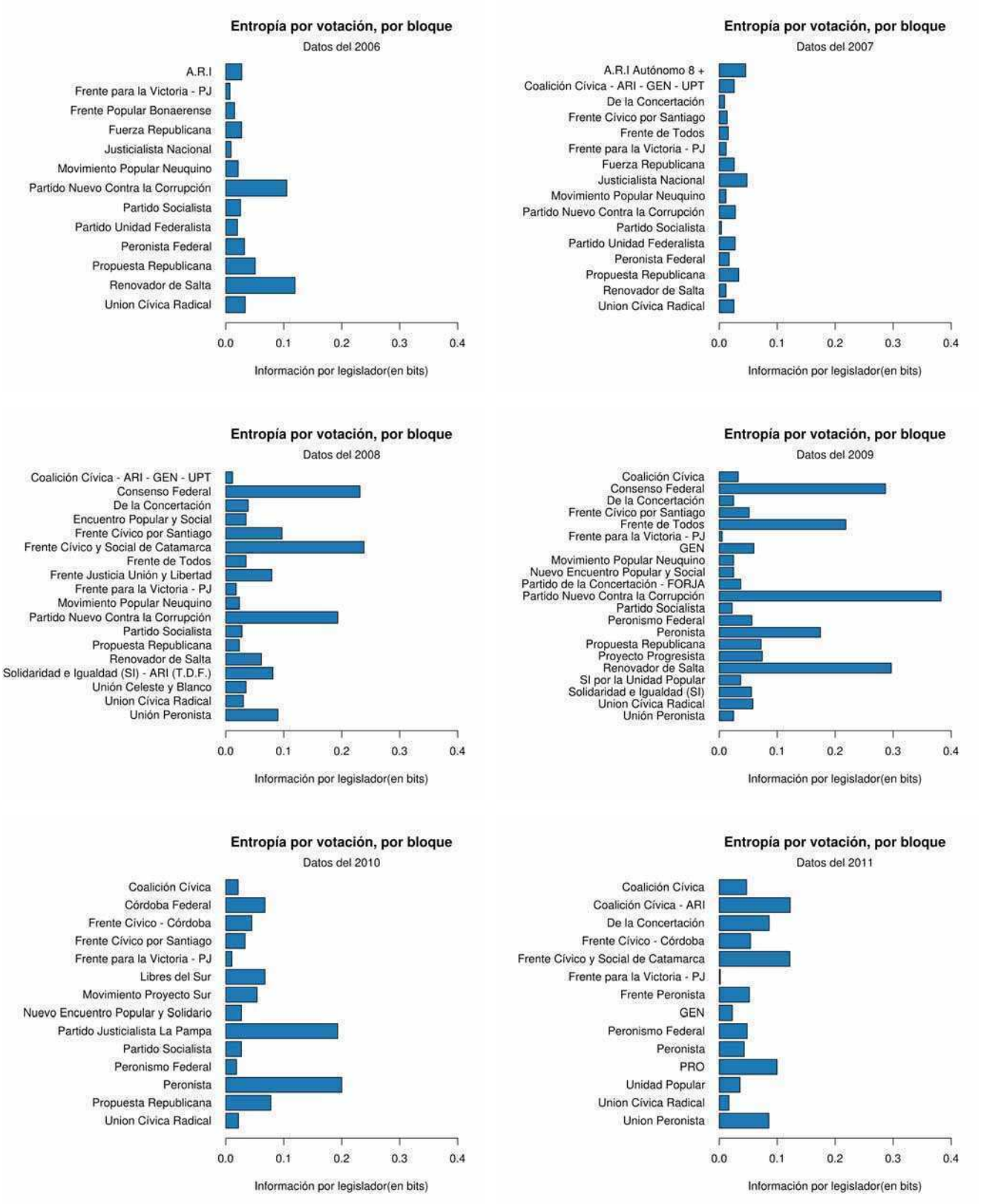

Figura 22: Bits de información por votación, por legislador, según el bloque.

Como se observa en la Figura 22, la cantidad de información por legislador por votación es notablemente baja, que indica que los diputados realmente votan en bloque, como era de esperarse. También es un resultado esperable que los bloques grandes son los que más parecen afectados por este fenómeno. Además, y un resultado novedoso, es que el efecto de este fenómeno varía considerablemente de año a año, en más de un orden de magnitud. Hay años en que los partidos alinean mejor a sus diputados consiguiendo una entropía notablemente baja, y otros años en que los diputados presentan una mayor independencia.

Dependencia entre bloques mediante análisis de entropía

Se repite el análisis del punto anterior, esta vez asignándole un voto a cada bloque, para obtener la cantidad de "bloques platónicos equivalentes" de toda la cámara. Al igual que para el análisis anterior, en caso de empate se considera que gana Afirmativo. Esto sólo es considerando un voto por bloque, y el resultado puede ser completamente distinto al resultado de la votación en el recinto, contando cada legislador.

El voto de cada bloque se calcula como la mayoría simple dentro de cada bloque. En el caso de que todo el bloque estuviera Ausente o votara por Abstención (caso no tan raro entre los bloques de una sola persona) nuevamente se considera que el voto va para la mayoría. Siendo que se le asigna un voto a cada bloque sea pequeño o grande, se espera que los bloques pequeños aporten la mayoría de la entropía de los resultados, así que se muestran varias corridas para cada año, quitando los bloques de menos de cierta cantidad variable de legisladores.

El código para este análisis se encuentra en `shannon/interbloque.py`.

En la Tabla 4 se muestran los resultados por año y por tamaño mínimo de bloque.

Año	Cantidad de bloques	Tamaño mínimo (en legisladores)	Bloques platónicos equivalentes
2006	33	1	5.252
	8	5	3.915
	5	11	3.222
	3	15	2.039
2007	48	1	5.519
	12	5	4.216
	5	11	3.15
	3	15	2.378
2008	38	1	5.604
	10	5	4.255
	6	9	3.6
	3	11	2.463
2009	47	1	5.844
	9	5	4.33
	5	9	3.395
	3	15	2.47
2010	36	1	5.048
	11	5	3.558
	5	9	2.786
	4	13	2.485
2011	50	1	4.23
	17	5	3.582
	9	9	2.966
	5	15	2.713

Tabla 4: Cantidad de entropía entre bloques para distintos tamaños de bloque, por año

Más allá de la baja cantidad de información en general por votación, es notable que no son los grupos pequeños los que aportan la mayoría de la entropía, ya que analizando únicamente los bloque mayoritarios obtenemos alrededor de la mitad de los bloques platónicos equivalentes, cada año. Los bloques más pequeños por más que en este análisis tienen un orden de magnitud más de votos (50 vs. 5 bloques por ejemplo en el 2011), apenas duplican la entropía total de la cámara. Es decir que, si bien

los bloques pequeños no están ligados formalmente a los bloques mayoritarios, tienden a aliarse a uno u otro de manera de que terminan aportando poca variedad en las votaciones.

Visualización de las votaciones con MDS y PCA

En el capítulo 9 de [Conway 2012] se propone un análisis interesante, que es visualizar las votaciones de los senadores de los Estados Unidos utilizando Multidimensional Scaling (MDS). Esta técnica toma como entrada una matriz de distancias y construye un mapeo a posiciones en un espacio de dimensiones especificadas que intenta conservar lo más posible las distancias originales, de manera que se pueda visualizar con facilidad.

Para aplicar la misma técnica a los datos de las votaciones nominales, se tomaron distancias euclídeas, luego de asignar a cada diputado una posición en un espacio N-dimensional, igual al utilizado en las secciones de *Agrupamiento Jerárquico* y *Agrupamiento mediante k-medias*. A cada voto Afirmativo se le asignó un valor de +1, a un voto Negativo el valor -1, y Ausentes y Abstenciones fueron imputados con 0.

Por otro lado, una técnica similar muy utilizada es el Análisis de Componentes Principales (PCA). Según [Cox 2001] existe una dualidad entre el MDS clásico métrico (también llamado Análisis de Coordenadas Principales o PCO) y PCA cuando la matriz de distancias está dada por distancias euclídeas.

Al igual que MDS, PCA permite visualizar un conjunto de puntos de alta dimensionalidad en un espacio de dimensionalidad menor, conservando las distancias relativas en la mayor medida posible. A diferencia de MDS, PCA no es sólo un método para visualizar los datos en un espacio de dimensionalidad inferior; permite decidir cuántas dimensiones son necesarias para capturar cierta cantidad de la variabilidad del sistema. Para realizar PCA es necesario tener los datos multivariados como puntos en el espacio de origen, no sólo la matriz de distancias entre los puntos. Si se dispone de esta información (como en el caso actual de las votaciones nominales), PCA permite visualizar vectores del espacio de origen proyectados en el espacio de destino.

Los resultados de correr MDS y PCA sobre las votaciones nominales del año 2009 se muestran en la Figura 23, restringido a los partidos mayoritarios.

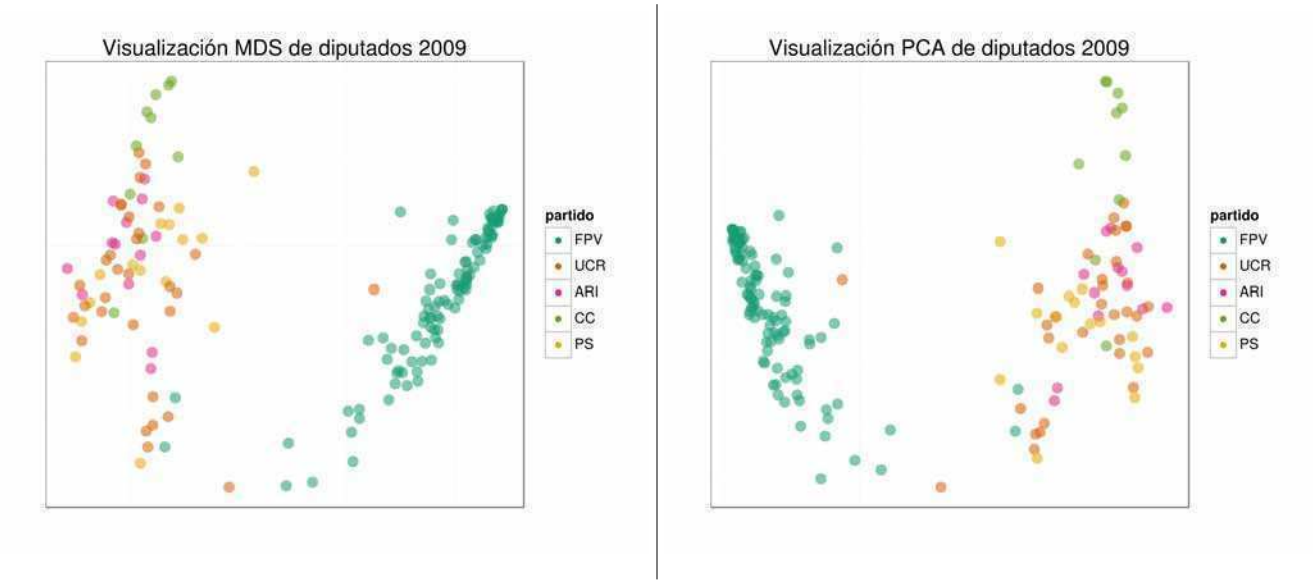

Figura 23: Visualización de las votaciones del 2009 mediante MDS y PCA

Es notable la similitud entre ambas proyecciones, y en ambas se hace evidente la brecha entre la forma de votar del oficialismo y la oposición. Por otro lado, utilizando PCA podemos analizar los autovalores generados, como se muestra en la Figura 24.

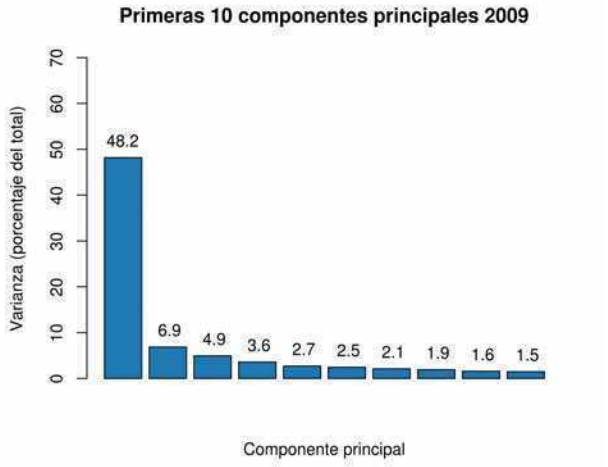

Figura 24: Autovalores generados con PCA para votaciones del 2009

Con estos autovalores se observa que, si bien ni siquiera las diez primeras componentes principales

llegan a capturar el 80% de la variabilidad del sistema, la primera componente principal es un orden de magnitud más significativa que las siguientes.

Finalmente, lo que nos permite el PCA es proyectar los vectores unitarios en el espacio de origen, en el conocido biplot de la Figura 25. Cada flecha allí representa una votación legislativa, y en particular el vector V104 representa la "Solicitud de tratamiento sobre tablas del expediente 5535-D-09" que se mencionó en la sección *Búsqueda de cliques en el grafo de la relación "concuerda con"*. Ésta es la única votación del año en la cual votar por Afirmativo indica fuertemente una postura de oposición. En la gran mayoría de las votaciones un voto Afirmativo es indicativo de oficialismo.

Algunas votaciones como la V49, "Exp. 152-S-08 (y otros) - O.D. 2109 - Vot. en Gral. y Part." acerca a los legisladores socialistas, ubicados en la parte superior del grupo. Efectivamente, el expediente 152-S-08, "Modificación del Código Civil y del Código de Comercio respecto a la mayoría de edad" fue un proyecto impulsado por un senador Socialista.

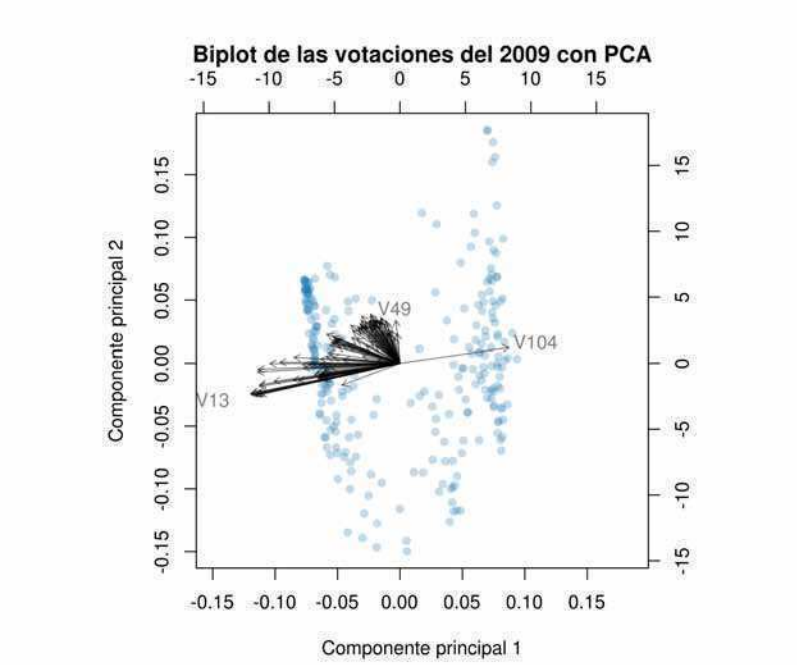

Figura 25: Biplot de legisladores y votaciones nominales en 2009.

Hasta aquí se han analizado las votaciones nominales de los legisladores únicamente. Sin duda el

resultado más útil de esta parte es el de los puntos ideales de los legisladores, y es un resultado que se incorporará en las partes siguientes. En la segunda parte de esta tesis se analizan los datos de los proyectos legislativos, pero es necesario recordar los resultados obtenidos en esta parte para enriquecer aquellos obtenidos más adelante.

Tanto la evolución histórica de los puntos ideales, como las técnicas de agrupamiento y el análisis de entropía en las votaciones proveen visualizaciones de la información que sería interesante poder monitorear en la actualidad y durante los años venideros, para saber si el comportamiento de un legislador varía de lo que los votantes esperaban, por ejemplo, o si aparecen quiebres dentro de un bloque o alianzas entre bloques antes de que estos sean explicitados o descubiertos por la prensa. Saber que hubo una variación en el comportamiento de un legislador durante el 2006 quizás no es tan interesante ni útil como saber que se produce el cambio apenas ocurre.

Finalmente, el último análisis de componentes principales parece indicar que habría alguna riqueza adicional en repetir el análisis utilizando dos dimensiones para el espacio de votaciones, para avanzar más allá de la dicotomía del oficialismo contra la oposición. En las siguientes dimensiones, incluso más allá de la segunda, deberían aparecer las verdaderas opiniones de los legisladores independientes de las órdenes del partido, o los principios del partido mismo independizado de una simple postura de apoyo o enfrentamiento al oficialismo.

Análisis de proyectos

Adquisición y análisis univariado

La página web de la Cámara de Diputados cuenta con un buscador[9] que provee detalles sobre todos los proyectos ingresados desde 1999.

Utilizando un script Python se adquieren secuencialmente todos los proyectos presentados en diputados durante los años 2006 al 2011, con los siguientes campos para cada proyecto:

- El tipo de proyecto. Se encuentran tres tipos de proyectos: de Resolución, de Declaración y de Ley.
- El número de expediente para el proyecto, que incluye el año en el que fue admitido.
- N° de trámite parlamentario en el que fue publicado el proyecto.
- Fecha en que fue presentado en mesa de entradas.
- URL para la página con detalles acerca del proyecto.
- El sumario del proyecto, un párrafo descriptivo o título.
- La lista de diputados que firmó el proyecto.
- La lista de comisiones en las que se discutió el proyecto.
- La lista de acciones tomadas en el transcurso del trámite del proyecto.
- El texto completo del proyecto, incluyendo fundamentos.

Estos datos se guardan no ya en un archivo de texto plano sino en una base de datos relacional PostgreSQL. El esquema entidad-relación del modelo es relativamente sencillo, y se muestra en la Figura 26.

9 http://www1.hcdn.gov.ar/proyectos_search/bp.asp

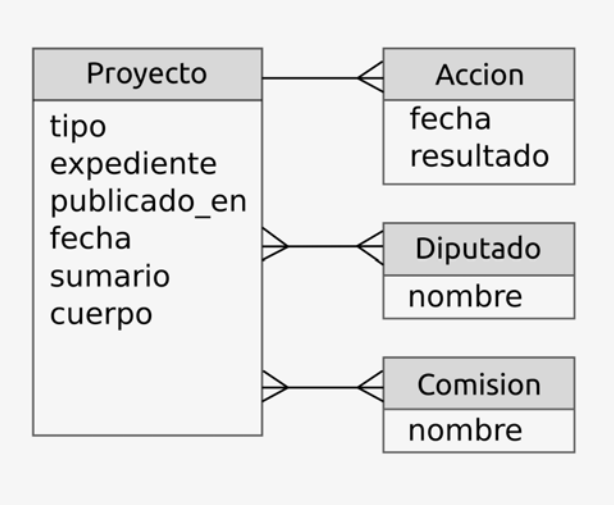

Figura 26: Diagrama del modelo de datos de proyectos

Los datos extraídos, incluyendo el DDL para generar las tablas utilizadas se encuentran adjuntos en el archivo `datos/proyectos.sql.gz`.

Así se obtuvieron un total de 32755 proyectos, de los tipos que se muestran en la Tabla 5.

Tipo de proyecto	N° de proyectos
Declaración	7324
Ley	9338
Resolución	16093

Tabla 5: Número de proyectos presentados según el tipo

La cantidad de proyectos presentados por año se muestra en la Tabla 6, y la distribución de proyectos presentados por fecha se grafica en la Figura 27.

Año	N° de proyectos
2006	5811
2007	4573
2008	5475
2009	4927
2010	7031
2011	4938

Tabla 6: Cantidad de proyectos presentados por año

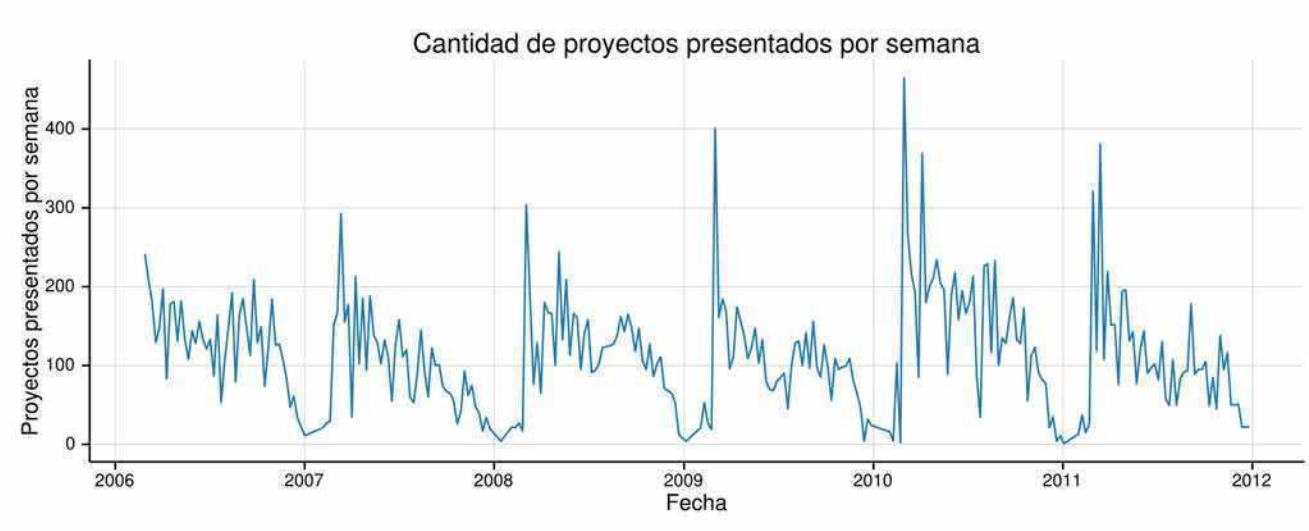

Figura 27: Distribución de proyectos presentados por fecha

Es interesante notar la estacionalidad de las labores legislativas, con un pico de proyectos presentados al inicio del período legislativo cada año, y un valle a fin de año correspondiente al cese de actividades y vacaciones.

Punto ideal del proyecto

Una herramienta útil para posteriores análisis sería tener un punto ideal del proyecto legislativo. Intuitivamente, legisladores oficialistas presentan proyectos oficialistas, y legisladores opositores presentan proyectos de oposición, así que se debería poder asociar el punto ideal del proyecto directamente al del legislador que lo presentó.

Un problema de este análisis es que algunos proyectos son presentados en conjunto por más de un

legislador, por lo que el punto ideal del proyecto no se obtiene directamente sino a través de alguna media. Para ver cuánto inciden estos proyectos presentados en conjunto en el análisis, se muestran la cantidad de firmantes por proyecto en la Figura 28.

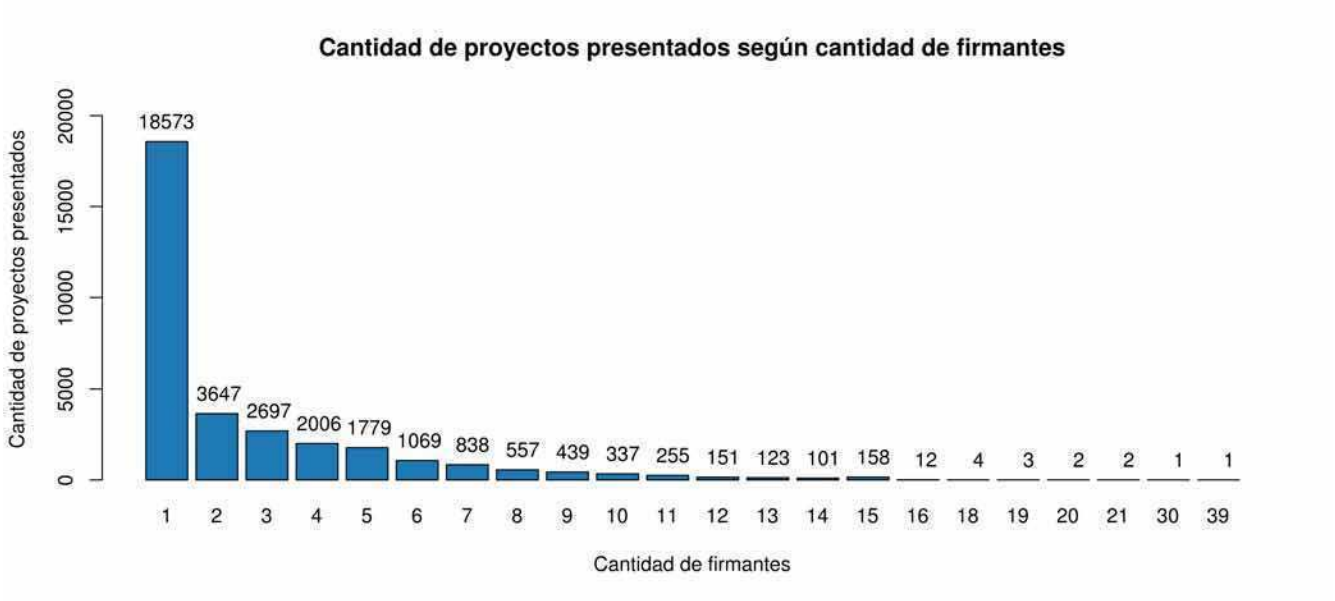

Figura 28: Cantidad de proyectos presentados por cantidad de firmantes

Como se observa, más de la mitad de los proyectos presentados son firmados por un único diputado, y apenas un 12% tiene más de 5 firmantes.

Por otro lado, en la Figura 29 se muestra la cantidad de proyectos presentados por dos o más legisladores según el desvío estándar del punto ideal de los firmantes. En apenas 334 proyectos, es decir menos del 4%, el desvío estándar es superior a 0.707, que sería equivalente a un proyecto firmado por dos legisladores con una diferencia en su punto ideal de 1.

Finalmente, un resultado obtenido por [Alemán 2009] es que existe una correlación entre los puntos ideales de los legisladores obtenidos de analizar las votaciones nominales y los que se obtienen de analizar los proyectos presentados en conjunto; por lo que se considera razonable considerar que el punto ideal de un proyecto es la media de los puntos ideales de los firmantes.

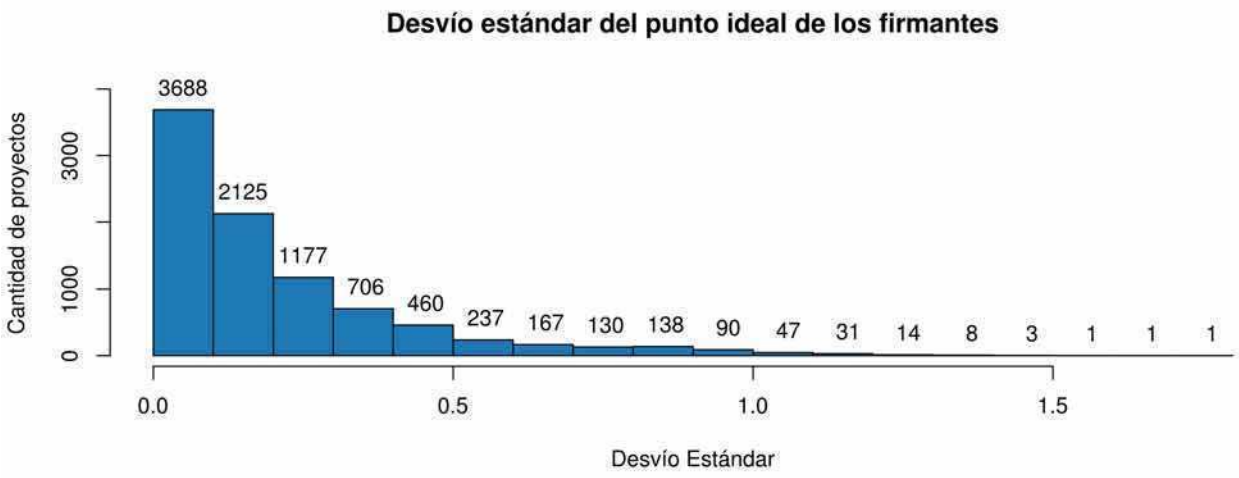

Figura 29: Cantidad de proyectos presentados, según el desvío estándar del punto ideal de los firmantes

Una vez obtenidos los puntos ideales de los proyectos de esta manera, se pueden graficar la cantidad de proyectos de distinto tipo presentados por punto ideal, como se muestra en la Figura 30.

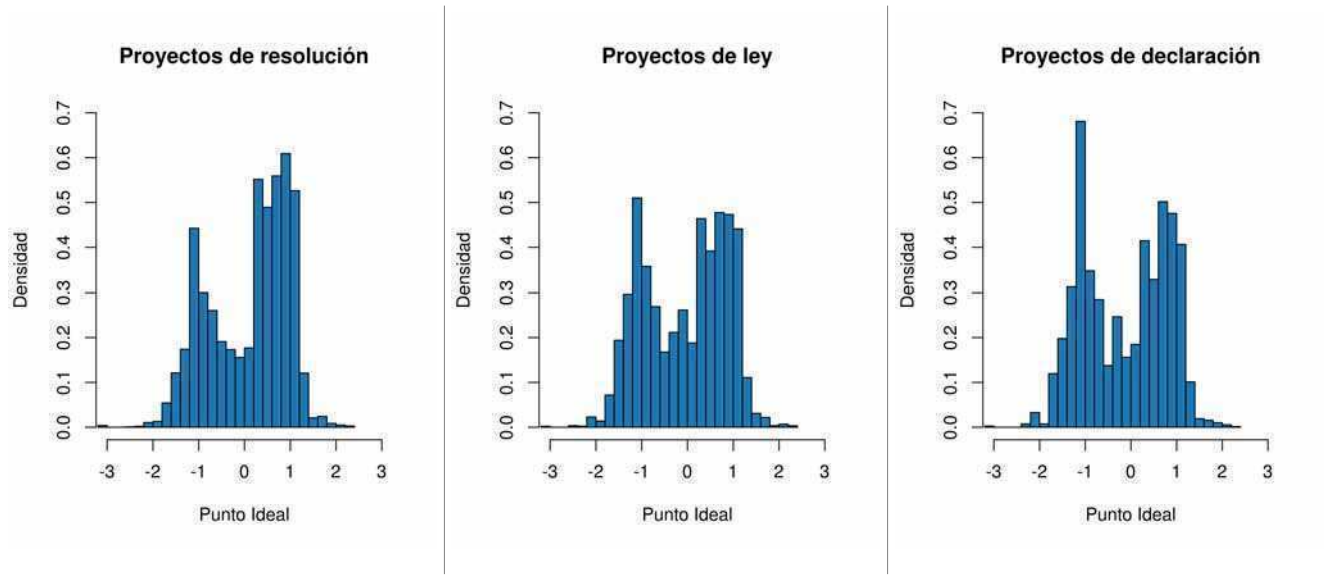

Figura 30: Histogramas de proyectos presentados según punto ideal y tipo

Se puede apreciar en los histogramas el mismo perfil bimodal que para los puntos ideales de los legisladores, por lo que en principio los diputados de distintas posturas legislativas presentan la misma

proporción de proyectos de distintos tipos.

Proyectos aprobados

Un dato de interés es cuáles proyectos fueron aprobados, y cuáles no. Surge una complicación cuando se intenta obtener este dato, ya que la gran mayoría de las veces la alternativa no es una pronunciación por la negativa, sino que el proyecto simplemente no es tratado, y el expediente permanece en los archivos, hasta que en algún momento es necesario considerar que el proyecto no va a ser aprobado nunca.

Según el reglamento de la Cámara de Diputados[10], los proyectos de resolución o declaración que no sean aprobados durante el año parlamentario en el que tuvieron entrada al cuerpo, se tendrán por caducados.

Los proyectos de ley que no sean sancionados en una de las dos Cámaras durante el mismo año o en el siguiente, también se tendrán por caducados. Si obtienen sanción en alguna de las dos Cámaras en el término indicado, el plazo se prorrogará por un año más.

El reglamento indica varias modificaciones y variantes, como los proyectos de códigos tratados con naciones extranjeras, o los proyectos enviados por el Ejecutivo sobre provisión de fondos para pagar los créditos contra la Nación, que nunca caducan.

En definitiva, hacia fines del 2012 existen algunos proyectos del 2011 o anteriores que podrían todavía llegar a ser aprobados, pero a fines prácticos para el presente trabajo se considera que todo proyecto que no se encuentra aprobado al momento de la extracción de datos se da por efectivamente desaprobado. Esto no quita que eventualmente puede volver a presentarse el mismo expediente en años posteriores para volver a tratar un tema, pero para el presente trabajo se consideran expedientes separados.

De esta manera se calculan los proyectos aprobados y desaprobados, y se muestran los resultados por año en la Tabla 7.

10 http://www1.hcdn.gov.ar/dependencias/dip/congreso/regladip.pdf Ver sección sobre caducidad de expedientes

Año	Aprobados	Desaprobados	Porcentaje de aprobación
2006	2209	3602	38.01%
2007	1731	2842	37.85%
2008	1919	3556	35.05%
2009	1065	3862	21.62%
2010	465	6556	6.61%
2011	155	4783	3.14%

Tabla 7: Cantidad de proyectos aprobados y desaprobados por año

Es notable el descenso de la tasa de aprobación de proyectos en los últimos años. Esto puede explicarse debido a la pérdida de mayoría del oficialismo en la Cámara baja, o por efecto del menor número de votaciones como se vio en la Tabla 1. Por otro lado estos dos puntos podrían a su vez estar relacionados.

Otro resultado interesante, si bien algo esperable, es el que se obtiene cuando se grafica la tasa de aprobación respecto del punto ideal del proyecto. Para eso, en la Figura 31 se agruparon los proyectos en centiles según el punto ideal.

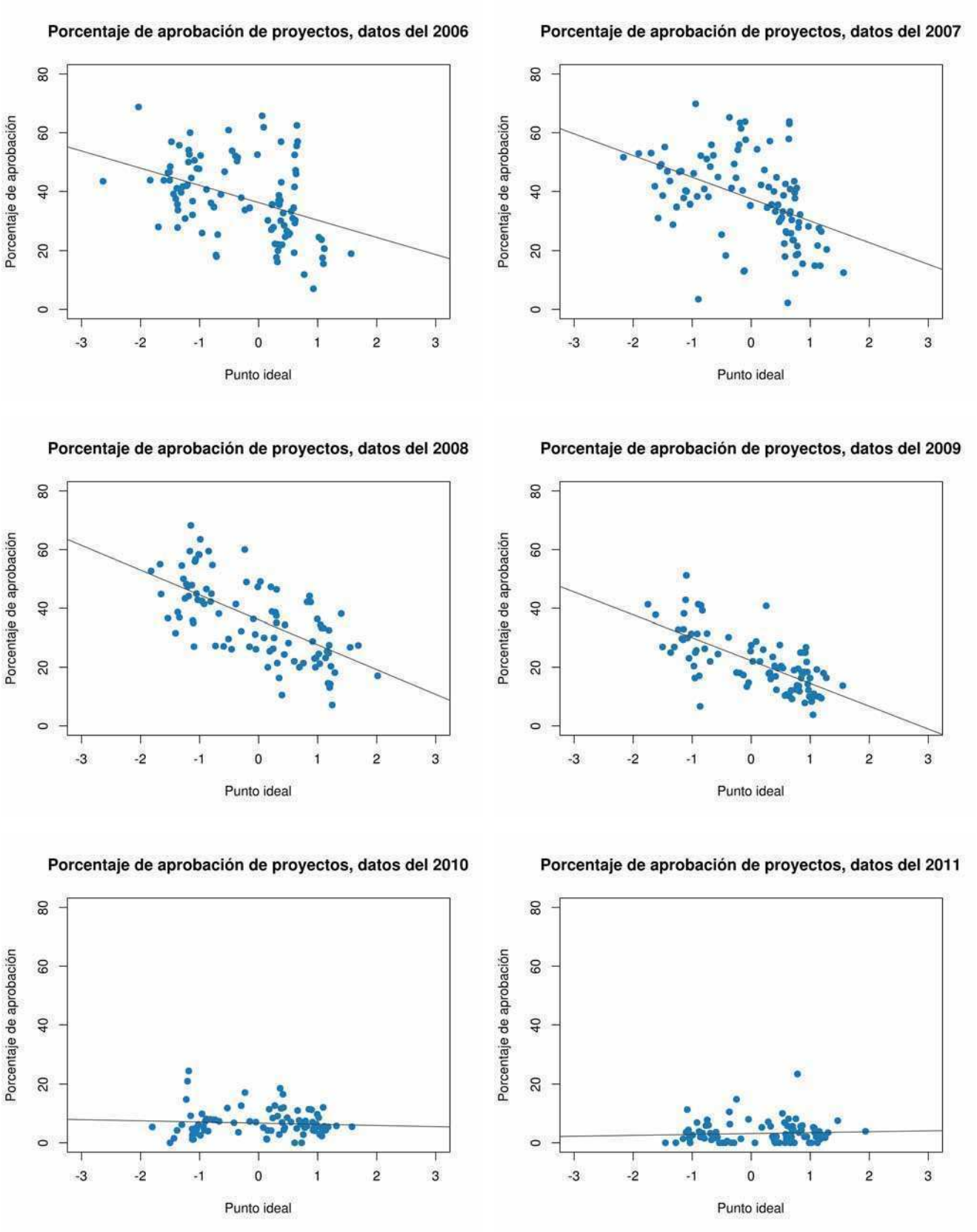

Figura 31: Porcentaje de aprobación de proyectos, según punto ideal, por año

En la Figura 31 cada punto corresponde a la tasa de aprobación de entre 45 y 70 proyectos, según el año. Sobre el gráfico de dispersión se sobreimpone la recta obtenida mediante una regresión lineal, utilizando la función `lm` de R.

En los últimos dos años, cuando el oficialismo pierde la mayoría en el congreso, cambia la pendiente de la regresión de manera notoria, a la vez que todos los sectores encuentran una tasa de aprobación más baja.

Para verificar que este cambio de pendiente es significativo, se realizaron una serie de pruebas, tomando la diferencia punto a punto entre centiles para cada par de años consecutivos, calculando una regresión lineal y tomando el t-valor de la pendiente de la recta obtenida, y la probabilidad de que datos tomados al azar podrían resultar en un t-valor que exceda en valor absoluto el obtenido (asumiendo la hipótesis nula de que la pendiente es cero). Los resultados de estas pruebas se muestran en la Tabla 8.

El código para estos tests se encuentra en `proyectos/t_tests.R`.

Años	Pendiente estimada	Desvío estándar	t-valor	P(>|t|)
2006 vs. 2007	-0.0140272	0.0215846	-0.650	0.517
2007 vs. 2008	-0.01337	0.01506	-0.888	0.377
2008 vs. 2009	0.02020	0.01346	1.5	0.137
2009 vs. 2010	0.070461	0.009986	7.056	2.46e-10
2010 vs. 2011	0.008163	0.007418	1.100	0.274

Tabla 8: T-tests de comparación de pendientes para años consecutivos

Como se esperaba, el único test que claramente rechaza la hipótesis nula es el de 2009 vs. 2010, por lo que el cambio en el comportamiento entre estos dos años no es solamente visiblemente notable, sino que es estadísticamente significativo.

Este análisis de los proyectos legislativos presentados brinda resultados de interés en sí mismos, por más que inicialmente se lo planteó simplemente como un nexo entre las votaciones nominales de los legisladores y los artículos periodísticos. El último resultado de la tasa de aprobación de los proyectos según su punto ideal sería valioso para analizar en otros años, según cambia la conformación y las distintas internas en la Cámara de Diputados.

Por otro lado, la asignación de un punto ideal a los proyectos es el paso clave para la parte siguiente: el análisis de la aparición en los medios de distintos proyectos legislativos, según la postura de los firmantes.

Análisis de artículos periodísticos

En esta tercera y última parte del trabajo se analizan artículos de los archivos de los diarios Clarín, La Nación y Página/12. Una intuición de la que se parte para este análisis es que los diarios Clarín y La Nación se encuentran sumamente alineados y en una postura opositora al gobierno. Mientras que el tercer diario, Página/12, durante los años analizados al menos, se encuentra en una postura alineada al gobierno. Asumiendo que éste es el caso se deberían poder encontrar similitudes entre Clarín y La Nación y diferencias entre estos y Página/12 en cuanto a los temas abordados, los términos utilizados, e incluso la atención que se le da a distintos proyectos y legisladores del oficialismo y la oposición.

Una vez completado un análisis preliminar del conjunto de datos se comenzará entonces por inspeccionar estas similitudes y diferencias entre los distintos medios adquiridos, para luego ver cómo se relacionan con los conjuntos de datos de las partes anteriores de esta tesis.

Adquisición

Se extraen y almacenan 915983 artículos de los archivos online de los diarios Clarín, La Nación y Página/12.

Para obtener estos datos nuevamente se recurre a un script Python. Es necesario utilizar código diferente para cada uno de los diarios, con la misma estructura básica pero ajustada al formato del archivo histórico de cada caso. El código para cada uno de estos scripts se encuentra en la carpeta `scraper_project/scraper/management/commands/` dentro de los archivos `scrape_clarin.py`, `scrape_nacion.py` y `scrape_pagina12.py` respectivamente. Como los datos son almacenados en una base de datos relacional y se pretende construir distintas visualizaciones, consultas y reportes sobre estos datos, se organiza el código dentro de una aplicación Django[11] que ya provee una buena organización y estructura para el código y los datos. Los campos almacenados de cada artículo se describe en el modelo de datos en `scraper_project/scraper/models.py` y cada comando para adquirir, extraer y procesar artículos es un comando Django que se ejecuta utilizando, por ejemplo:

```
$ python manage.py scrape_nacion
```

Este archivo `manage.py` se encuentra en `scraper_project/` y es el que permite realizar todas las operaciones relacionadas con el proyecto Django, desde inspeccionar la base de datos hasta correr el servidor web que provee las vistas y reportes.

11 https://www.djangoproject.com/

Para cada artículo se almacenó el html crudo, y además se extrajeron los siguientes campos:

- Título.
- Cuerpo del artículo.
- Volanta, el texto que va por encima del título, en tamaño menor, que introduce o adelanta el tema que se desarrolla en el artículo.
- Bajada (o copete), el texto que va por debajo del título y lo amplía, sintetizando la noticia en un párrafo.
- Sección del diario en el que apareció la nota.
- Fecha de publicación.

Todos los artículos extraídos se encuentran disponibles en el archivo `datos/notas.sql.gz`.

Un problema encontrado fue que el diario Clarín cambió el formato de su archivo histórico a mediados del 2010, y en el nuevo formato se encuentran disponibles muchos menos artículos que en el formato anterior. Se deberá tener esto en cuenta cuando se interpreten los resultados que involucren este diario en los dos últimos años.

La cantidad de artículos total por diario y por año se muestra en la Tabla 9, y la distribución de artículos por fecha se muestra en la Figura 32.

Año	La Nación	Página/12	Clarín
2006	77408	39454	44192
2007	78592	54985	40984
2008	77993	50216	36508
2009	73208	52096	35071
2010	63588	58805	19724
2011	58835	47744	6580
Total	**429624**	**303300**	**183059**

Tabla 9: Cantidad de artículos por diario y por año

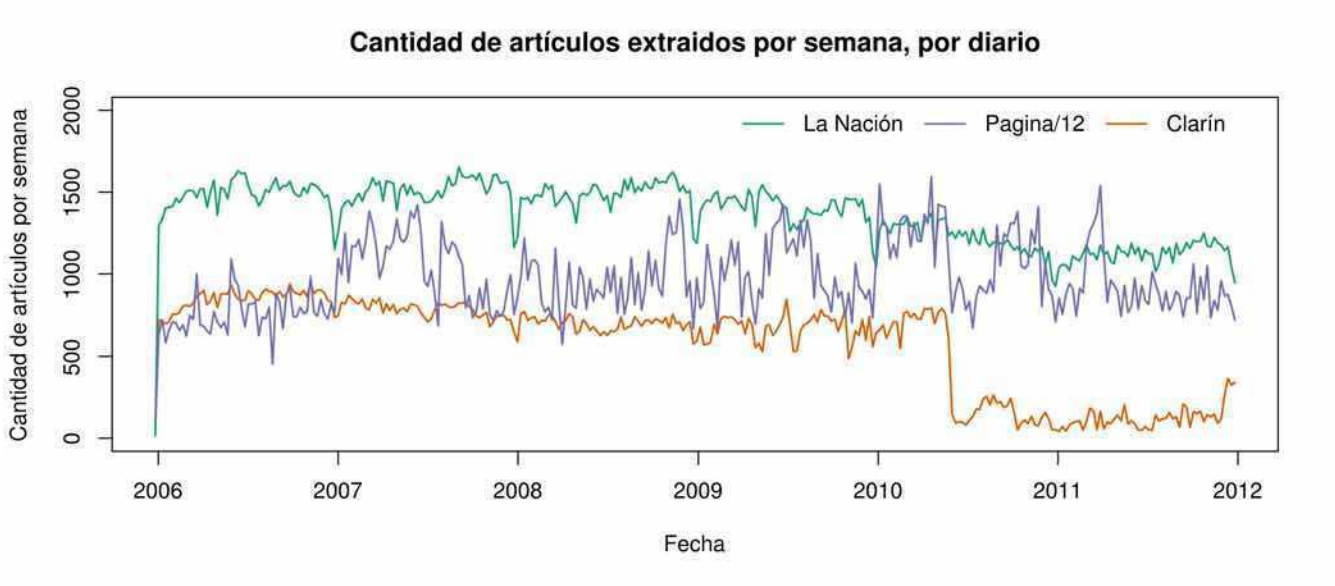

Figura 32: Artículos extraídos por semana, por diario

La sección de cada artículo fue obtenida del texto de los artículos mismos, según fue publicado en cada medio. Esto agrega una pequeña complejidad ya que los distintos diarios tienen distintos nombres para cada sección. La composición de artículos por sección para cada diario se muestra en la Figura 33.

Así, los temas legislativos se tratan en la sección *Política* del diario La Nación, que se asemeja a la sección *El País* o bien *Sociedad* de Clarín, que coincide con las secciones de Página/12. Algunos temas puntuales pueden tratarse en las secciones de *Economía* de La Nación o Página/12, pero Clarín no publica artículos en esta sección. Clarín publica el suplemento iEco que es una revista especializada en economía, pero no se extrajeron artículos de este suplemento.

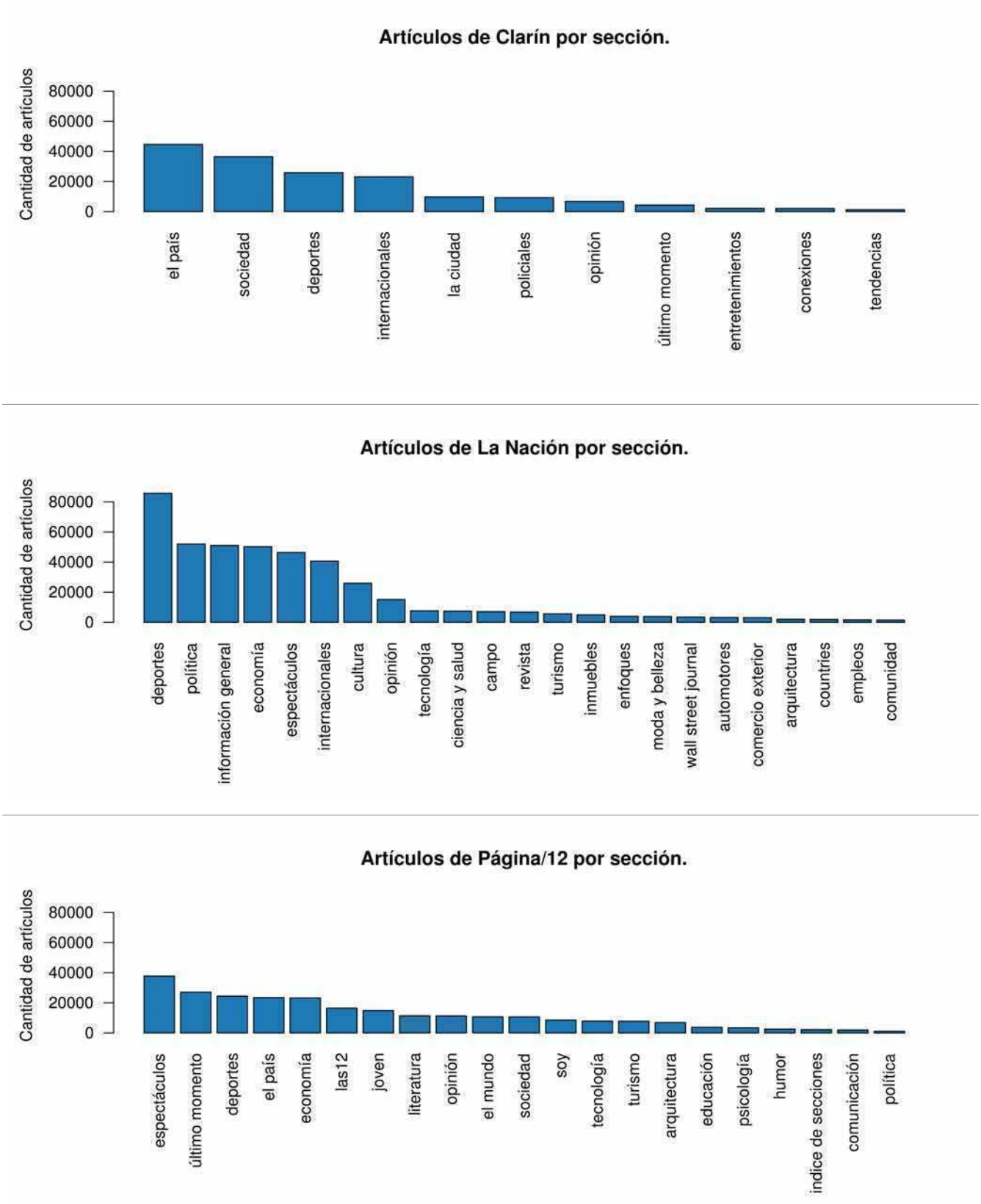

Figura 33: Cantidad de artículos por sección, por diario

Preprocesamiento y lematización

Al trabajar con el texto de los artículos, se realiza un primer preprocesamiento que consiste en remover las entidades html, normalizar el espaciado y convertir todo a codificación utf-8.

Luego, para algunas tareas de comparación de documentos, se realizó un paso adicional de lematización. La lematización es el proceso mediante el cual se asigna a cada palabra en un texto el lema correspondiente, donde un lema es la palabra que por convenio se acepta como representante de todas las formas flexionadas. Las formas flexionadas de un verbo serían todas las conjugaciones en todos los tiempos verbales, por ejemplo, donde por convención se acepta que el lema es el verbo en infinitivo. Las formas flexionadas de un sustantivo serían el plural, el diminutivo, y demás variantes del sustantivo, que a su vez se toma como el lema.

La mayoría de las veces la lematización requiere de un análisis sintáctico de la oración, ya que el simple análisis morfológico de las palabras puede ser ambiguo, y es necesario tomar cada palabra en el contexto de la oración para desambiguar. Por ejemplo, la palabra *sala* puede referirse a un verbo conjugado (si se dijera "Juan *sala* la comida"), o a un sustantivo ("Esta *sala* es amplia"). En el primer caso el lema correspondiente sería *salar* (el verbo en infinitivo), mientras que en el segundo sería *sala* (el sustantivo mismo).

La alternativa que muchas veces se considera a la lematización es el *stemming*, que consiste en remover los sufijos de las palabras para obtener una raíz común a todas las formas flexionadas. Si bien esta técnica es utilizada con relativo éxito en el idioma inglés, y en teoría debería funcionar con verbos regulares y sustantivos que pluralizan de distintas maneras, los resultados obtenidos en castellano a menudo son pobres [Honrado 2000].

Afortunadamente existen librerías que realizan el proceso de lematización de manera automatizada, especializadas al idioma castellano. La librería utilizada para el presente trabajo es FreeLing[12], desarrollada por Lluís Padró y el Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla, de la Universidad Politécnica de Cataluña[13]. Esta librería posee varios módulos de análisis sintáctico configurables por separado [Padró 2010], además de proveer bindings para el lenguaje Python.

En la Tabla 10 se muestran algunos ejemplos de fragmentos de textos lematizados utilizando esta librería.

12 http://nlp.lsi.upc.edu/freeling/
13 http://www.talp.upc.edu/

Texto original	Texto lematizado
Juan ama a Ana.	juan amar a ana .
La ama de llaves trajo los abrigos.	el ama de llave traer el abrigo .
Mi amigo Juan Traje trajo el traje para la boda.	mi amigo juan_traje traer el traje para el boda .
Un total de diez activistas de Greenpeace fueron detenidos ayer por la Prefectura uruguaya cuando realizaban una protesta contra la instalación de las dos papeleras en Fray Bentos y fueron liberados pocas horas más tarde.	uno total de activista de greenpeace ser detener ayer por el prefectura uruguayo cuando realizar uno protesta contra el instalación de el papelera en fray_bentos ser liberar poco hora más tarde .

Tabla 10: Ejemplos de textos lematizados.

Selección del vocabulario

Para los análisis de texto un problema recurrente es la selección del vocabulario a considerar. Por poner un ejemplo, si tomamos los artículos periodísticos del 2009 lematizados, obtenemos un corpus de 111810 documentos que cuenta con un vocabulario de 402765 lemas. Esto es una vez lematizado, es decir que muchas de las formas flexionadas han sido removidas, reduciendo el vocabulario inicial. De este vocabulario inicial, 208911 lemas (más del 50%) son utilizadas una sola vez en todo el corpus. La Figura 34 lista una muestra aleatoria de 20 términos con la cantidad de utilizaciones en el 2009.

Lema	Ocurrencias	Lema	Ocurrencias
enestese	1	armonicista	3
ganglio	31	marcelo_weissel	9
marramao	1	worms	1
bárbara_rosen	12	anuka	15
rociar	250	mujermadrid	1
coach	288	naím	9
hibernia	2	diana_maggi	8
quiniela	64	elcapo	1
dirgido	2	akhenatón	1
travis_payne	1	frank_rhodes	1

Figura 34: Algunos lemas del corpus de 2009, con su cantidad de ocurrencias

Como trabajar con el vocabulario completo y crudo tal cual aparece en el corpus resulta computacionalmente caro y arroja resultados mediocres, en una primera implementación se utiliza una lista arbitraria de stop words, y se eliminan los términos con una sola ocurrencia.

Trabajando con vocabularios reducidos de esta manera todavía los resultados son poco claros, sobre todo para la generación de tópicos con LDA donde se dificulta la rotulación de los mismos debido a que los términos más ponderados en cada tópico parecen tener poca relación entre sí.

Se busca entonces una manera mejor de definir el vocabulario a utilizar.

Una herramienta útil para esto es el *tf-idf*, una medida de lo importante que es un término para un documento dentro de un corpus. Éste es el producto de dos estadísticas, la frecuencia del término (*tf* o term frequency) y la inversa de la frecuencia de documento (*idf* o inverse document frequency). Dado un término *t* y un documento *d* dentro de un corpus *D*, la frecuencia del término *tf(t, d)* está dado simplemente por la cantidad de ocurrencias del término dentro del documento. La inversa de la frecuencia de documento a su vez está dada por

$$idf(t, D) = \log \frac{|D|}{|d \in D : t \in d|}$$

donde |*D*| es la cardinalidad de *D*, es decir la cantidad de documentos en el corpus, y $|d \in D : t \in d|$ es la cantidad de documentos en el corpus que contienen el término *t* al menos una vez.

Entonces se define el *tf-idf* como

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

Esta estadística resulta sumamente útil ya que captura la importancia de un término para un documento. Se puede representar un documento como “vector de tf-idfs” donde cada documento se representa por un vector con un elemento por cada término en el vocabulario, y cada elemento es el tf-idf del término dentro del documento. En algunos casos esta representación resulta más adecuada que la representación más habitual de “bolsa de palabras”; en este trabajo se utiliza la representación de “vector de tf-idfs” en las secciones de *Comparación de posturas legislativas: Recalde, Macaluse y Quiroz*, y en algunos clasificadores preliminares de la sección de *Correspondencia entre proyectos y artículos*.

Volviendo al problema de la selección de vocabulario, se prueban tres alternativas:

- La sugerencia de [Blei 2009] que es utilizar un vocabulario de tamaño fijo, consistente en los N términos de mayor tf-idf a nivel corpus, una variante del tf-idf que cuenta la frecuencia del término en todo el corpus, para independizarse de cualquier documento en particular:

$$tfidf_{Global}(t,D) = tf(t,D) \cdot idf(t,D)$$

- Obtener un vocabulario de tamaño fijo consistente en los N términos de mayor tf-idf promedio en todo el corpus:

$$tfidf_{Avg}(t,D) = \frac{\sum_{d \in D} tfidf(t,d,D)}{|D|}$$

  Esta variante es equivalente a la anterior salvo por un factor de escala de $|D|$, por lo que el vocabulario resultante es idéntico.

- Obtener un vocabulario de tamaño fijo consistente en los N términos de mayor tf-idf máximo en todo el corpus:

$$tfidf_{Max}(t,D) = max(tfidf(t,d,D) : d \in D)$$

En la Figura 35 se muestra la primera parte (en orden alfabético) del vocabulario de 5000 lemas generado por estas dos estrategias con el corpus de los artículos de todo el 2009. Los lemas con fondo gris son los términos en común entre los dos vocabularios, mientras que los términos con fondo rojo son términos que a simple vista sería preferible que no estuvieran.

Max	**Avg / Global**
aa aaavyt aaba aaccpp aacs aacumar aado aae aaea aaeeh aaetav aaofp aapa aapbp aapoyo aapresid ab abbsa abc abigeato abogacía abogado aborigen aborto abra_pampa abreviar abriendo_surcos abrogar abs absorbente abuelas abuelo abuso academia acampante accae accidente acciones_clase aceite	abajo abandonar abandono abarcar abastecer abastecimiento aberrante abocar abogado abonar abordaje abordar a_bordo aborigen aborto abril abrir absolutamente absoluto absorber abstener abuelo abundar abusivo abuso acá acabar academia académico acaecer a_cambio_de a_cargo_de acarrear a_causa_de acceder accesibilidad accesible acceso accesorio accidente acción accionar accionario accionista aceite

Figura 35: Vocabularios generados con distintos criterios de selección

La estrategia del máximo tf-idf (la lista de la izquierda) contiene todos términos menos frecuentes, abundan las siglas y los nombres propios, y llega al punto de tener lo que parecen ser errores tipográficos comunes, que sería mejor eliminar. La estrategia de tomar el tf-idf global o promedio (la lista de la derecha) contiene términos más comunes, al punto de contener lemas que probablemente deberían ser considerados stop words.

Se consideró la alternativa de formar un vocabulario de tamaño *N* encontrando primero el *M* tal que la

intersección de los vocabularios de tamaño *M* generados por ambas estrategias sea de tamaño *N*. El vocabulario resultante de esta manera es interesante pero todavía contiene stop words comunes (por ejemplo para un vocabulario de 2000 términos sobre el corpus del 2009 ya aparecen los lemas *el*, *lo*, y *ello*), por lo que se prefirió conservar la estrategia más sencilla del tf-idf global, y filtrar el vocabulario con una lista de stop words. Ésta es la estrategia utilizada para todos los puntos donde interviene un vocabulario generado sobre un corpus de texto, salvo que se indique lo contrario. La lista completa de palabras eliminadas del análisis se incluye en el Apéndice B.

Análisis de tópicos con LDA

Se comienza entonces con un análisis cualitativo de las diferencias entre los tópicos abordados por cada uno de los tres diarios analizados.

El Latent Dirichlet Allocation (LDA) es un modelo generativo que supone que cada documento es producto de una mezcla de tópicos, que son a su vez distribuciones de probabilidad sobre el vocabulario completo con el que se trabaja.

En [Blei 2011] se explica este modelo generativo describiendo el proceso que se utiliza para generar un documento. Se asume primero que existe alguna cantidad de tópicos dados por una distribución de probabilidad sobre un vocabulario fijo. Por ejemplo, el tópico de *Genética* tendrá una mayor probabilidad asignada a términos de genética, y una menor probabilidad para términos relacionados con carpintería. Luego, cada documento se genera en un proceso de dos etapas:

1. Se elige la distribución de tópicos para el documento de manera aleatoria.
2. Para cada palabra, se elige primero el tópico al que pertenece, de acuerdo con la distribución elegida en el paso 1., y luego se elige una palabra al azar de acuerdo con la distribución del tópico.

Utilizando la librería `gensim`[14] de Python se estimaron los tópicos de cada diario tomando los artículos de los tres diarios del 2006. Se fijaron 5 tópicos, y el vocabulario fue tomado por los 5000 términos más frecuentes de cada diario, utilizando los textos lematizados.

Para cada diario, en la Tabla 11 se muestran los 10 términos más probables para cada tópico, junto con un nombre para el tópico elegido de manera manual.

14 http://radimrehurek.com/gensim/

Clarín				
Deportes	**Política**	**Internacionales**	**Economía**	**Policiales**
equipo	presidente	juez	gobierno	policía
partido	gobierno	caso	empresa	casa
jugar	político	causa	precio	persona
ganar	ministro	juicio	sector	mujer
jugador	kirchner	derecho	aumento	fuente
punto	jefe	corte	medida	hombre
final	elección	persona	ley	hijo
boca	gobernador	presentar	mercado	chico
gol	candidato	justicia	proyecto	morir
local	estado	pedir	subir	joven
club	militar	tribunal	público	víctima
fútbol	fuerza	obra	servicio	informar
La Nación				
Deportes	**Política**	**Espectáculos**	**Economía**	**Sociedad / otros**
equipo	gobierno	obra	empresa	hijo
jugar	presidente	informes	precio	casa
partido	político	entrada	mercado	mujer
ganar	ministro	música	sector	lugar
jugador	policía	presentar	gobierno	chico
final	juez	libre	servicio	hombre
punto	ley	director	us	familia
boca	jefe	público	producto	persona
club	persona	grupo	proyecto	foto
torneo	estado	trabajo	inversión	volver
gol	pedir	programa	sistema	salir
próximo	kirchner	libro	valor	padre
Página/12				
Deportes	**Política**	**Espectáculos**	**Cultura**	**Otros / opinión**
equipo	gobierno	dirección	mujer	obra
partido	presidente	trasnoche	libro	ciudad
jugar	político	sábado	escribir	lugar
jugador	empresa	viernes	historia	artista
ganar	sector	juez	pensar	casa
boca	proyecto	presentar	disco	historia
gol	nacional	domingo	momento	arte
salir	público	causa	película	siglo
final	provincia	atp	chico	abrir
terminar	estado	tel	hijo	buenos_aires
lugar	ministro	policía	hombre	presentar
punto	política	entrada	propio	libro

Tabla 11: 5 tópicos generados con LDA para los distintos diarios, datos del 2006

El código utilizado para generar estas listas de tópicos por diario se encuentra en el archivo `exploratorio_articulos/topics_2006.py`.

Estos resultados son de interés exploratorio, y de naturaleza cualitativa; en las siguientes dos secciones se intenta cuantificar la similitud o proximidad entre los tres medios analizados en base a los términos y tópicos utilizados. A los fines de este trabajo, los resultados de esta sección están abiertos a las conclusiones del lector, pero resulta de interés en principio que, con este procedimiento de selección de vocabulario y generación de tópicos se marcan claramente las principales secciones de cada diario. Sobre los tópicos generados, resulta de interés por ejemplo que Clarín no marca fuertemente un tópico de espectáculos, como sí lo hacen los otros dos diarios. Además, generando los tópicos de esta manera Clarín toma términos de un tópico que parece ser Economía, por más que no publica artículos bajo esta sección (cabe recordar que no se tomaron notas de la revista iEco), mientras que Página/12 no tiene un tópico que sea claramente de temas económicos, por más que tenga más de 20000 artículos publicados bajo esta sección.

Una observación es que el proceso de estimación de la distribución de tópicos de cada documento y de los tópicos mismos es no determinista como el proceso de generación de documentos en sí, por lo que dos corridas de LDA pueden resultar en tópicos notablemente distintos, dado el mismo corpus de texto.

Clasificación de origen en base al cuerpo del artículo

Para saber qué tan distinguible es el contenido de los artículos de los distintos diarios se procede a entrenar un clasificador que detecte el diario de origen en base exclusivamente al cuerpo lematizado del artículo.

Para esto se utiliza la librería Scikit Learn[15] para Python que dispone de varios mecanismos de clasificación listos para usar. Se trabaja con 6 meses de datos de los tres diarios, de enero a junio de 2006 inclusive.

Se entrenaron y evaluaron cuatro clasificadores disponibles con la librería:

- Un clasificador Naive-Bayes multinomial, como se describe en [Rennie 2003].
- Un perceptrón [Rojas 1996], con velocidad de aprendizaje $\alpha = 0.0001$ y 50 iteraciones. Como el clasificador básico distingue sólo dos clases se utiliza una estrategia de Uno Contra Todos

15 http://scikit-learn.org/

(OVA, por sus siglas en inglés), que consta de tres clasificadores en paralelo donde cada uno distingue una de las clases deseadas.

- Un clasificador de descenso por gradiente estocástico similar al perceptrón, pero con una velocidad de aprendizaje que disminuye de manera inversamente proporcional al número de iteración. Al igual que para el perceptrón se utiliza una estrategia OVA para implementar el clasificador multiclase.

- Una máquina de vectores de soporte multiclase con un kernel lineal, implementado con liblinear [Fan 2008].

En cada caso se entrena el clasificador con dos tercios de los artículos disponibles tomados al azar, y se testea con el tercio restante. El código utilizado en esta sección se encuentra en la carpeta `origin_classifier/`.

Inicialmente se realiza una corrida con la misma lista de stop words utilizada para la generación de tópicos con LDA, obteniendo las performances de clasificación que se muestran en la Figura 36. Aquí el score F_1 es la media armónica de la precisión y el recall promedio, es decir

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Figura 36: Performance de clasificación mediante distintas técnicas

El único clasificador que obtiene resultados notablemente más bajos es el Naive-Bayes, mientras que los otros obtienen puntajes similares, con la máquina de vectores de soporte levemente por encima del resto, con un score F_1 de 81%. En todos los casos el desempeño es mejor para La Nación, como se

muestra a continuación para la máquina de vectores de soporte:

```
              precision    recall  f1-score   support

   lanacion       0.84      0.88      0.86     12702
     clarin       0.74      0.72      0.73      6589
   pagina12       0.81      0.72      0.76      4383

avg / total       0.81      0.81      0.81     23674
```

Este resultado es esperable, ya que La Nación es la clase mayoritaria, con más muestras que las otras dos clases juntas.

En la Figura 37 se muestran los 10 términos que más peso tuvieron para cada clase, según la técnica utilizada.

Naive-Bayes			**Perceptrón**		
La Nación	**Clarín**	**Página/12**	**La Nación**	**Clarín**	**Página/12**
mañana	ocurrir	hombre	dpa	taiwán	aparte
recibir	local	social	fibertel	george_bush	tengo
informar	mantener	empezar	caricaturar	avenida_brasil	frente_amplio
local	volver	cierto	agencia_dyn	arrancar	atlético_madrid
esperar	grupo	estado	dyn	anchorena	lucas_livchits
caso	momento	escribir	berlin	súper	esta
la_nacion	político	terminar	cotejo	filme	cambios
público	jefe	punto	rojos	aerolíneas	pib
comenzar	juez	presidente	cordoba	apego	santafesino
lugar	hombre	abrir	agencias_afp	redmond	casa_gris
Clasificador SGD			**SVC Lineal**		
La Nación	**Clarín**	**Página/12**	**La Nación**	**Clarín**	**Página/12**
en_favor_de	kaiser	concejal	rueda	anotador	pib
us	madrugada	viejo	agencia_dyn	abu_mazen	recomer
opinión	tirar	guitarra	santafecino	de_villepin	transnacional
junto_con	a_favor_de	binner	por_medio_de	básquet	esta
cotejo	barsa	ciencia	censos	kaiser	aparte
paris	afuera	santa_fe	agencia_télam	george_bush	ud
respecto_de	papelera	cambios	jerusalen	nene	remarcar
agencias_dyn	mañana	remarcar	berlin	súper	casa_gris
seleccionar	juzgado_nº	disco	agencias_télam	ronda	lt
rueda	registrar	jueves	cotejo	cromañón	cambios

Figura 37: Términos de mayor peso según clase y clasificador

En varios casos los términos que más pesan para discriminar los distintos medios contienen lemas que nombran a los diarios mismos, y a las agencias de noticias que cada una utiliza (AFP, EFE, Télam, Reuters, DyN). Esto es de esperar, ya que un artículo que menciona la palabra Clarín tiene mayor probabilidad de haber sido publicado en el diario Clarín mismo. Como el objetivo es saber qué tan distinguibles son los medios por el vocabulario utilizado en los artículos, y los temas tratados, se repite la corrida agregando estos términos a la lista de stop words para este punto.

Al analizar los resultados es claro que varios de los términos más distintivos para los clasificadores ahora son distintas maneras de deletrear nombres, como “Shiitas” vs. “Chiitas”, o “Al Quaeda” vs. “Al Quaida”, incluso “Cromañón” vs. “Cromagnon”. Se realiza una corrida más, eliminando además estas variantes.

Finalmente, al analizar estos nuevos resultados se observa que los términos más distintivos para el diario Página/12 incluyen para varios clasificadores términos relativos a Rosario, como “santafesino”, “rosario” o “canalla”, que se deben a la inclusión de la edición Rosario/12 como una sección más del diario. Se repite la corrida, esta vez sin utilizar los artículos de esta sección.

Los scores F1 obtenidos para las distintas corridas con las distintas técnicas se muestran en la Tabla 12.

Datos	Naive-Bayes	Perceptrón	Clasificador SGD	SVC Lineal
Lista de stopwords	0.6787	0.7852	0.7527	0.8084
Stopwords con agencias	0.6557	0.7437	0.7089	0.7684
Stopwords con agencias y diferencias ortográficas	0.6496	0.7361	0.7081	0.7681
Todo lo anterior, sin Rosario/12	0.6492	0.7405	0.7035	0.7691

Tabla 12: Score F1 para distintas técnicas y distintos filtros de datos

Se observa en los resultados, como era de esperarse, que cada vez que se quita parte de los datos la performance de los clasificadores baja sutilmente, aunque la performance relativa de las distintas técnicas permanece igual. Naive-Bayes muestra consistentemente el peor rendimiento, y el SVC lineal queda siempre levemente por encima del resto. En todos los casos la disminución de la efectividad por la restricción del vocabulario y los artículos es de un 3%.

La lista de términos que más pesan para distinguir cada diario en el último caso se muestra en la Figura 38. Debe notarse que ahora la lista contiene palabras que podrían bien aparecer en uno u otro medio por igual, y aún así la efectividad del mejor clasificador (el SVC lineal) es superior al 75%.

Naive-Bayes			**Perceptrón**		
La Nación	**Clarín**	**Página/12**	**La Nación**	**Clarín**	**Página/12**
mantener	detener	cambio	fv	ruralistas	aparte
mañana	punto	hijo	informante	patrimonial	había
informar	ciudad	grande	comfuentes	latinoamérica	soy
esperar	volver	idea	londres	bebé	alca
caso	mantener	grupo	caricaturar	abu_mazen	eso
local	momento	trabajar	por_medio_de	rondó	asimetría
recibir	grupo	hombre	censos	aberrante	atlético_madrid
público	jefe	presentar	moscu	súper	rosada
comenzar	juez	presidente	tokio	niembro	cambios
lugar	jugar	punto	miembro_de_gda	anotador	básquetbol
Clasificador SGD			**SVC Lineal**		
La Nación	**Clarín**	**Página/12**	**La Nación**	**Clarín**	**Página/12**
aviario	asaltante	película	match	kiosco	hemos
poder_ejecutivo	oriente_medio	yomepregunto	repsol_ypf	aerolíneas	remarcar
automóvil	enseguida	recién	pues	abu_mazen	básquetbol
berlin	madrugada	incidencia	santafecino	enseguida	cambios
situar	estadounidense	mito	caricaturar	beijing	kuhn
londres	derrotar	bc	informante	niembro	eso
seleccionar	online	tenemos	rojos	imprimir	pib
us	afuera	traducción	hotmail	de_villepin	ud
doctor	anotador	esta	rueda	américa_latina	latinoamérica
cotejo	chances	ud	aviario	george_bush	lucas_livchits

Figura 38: Términos de mayor peso, eliminando lemas de agencias y medios, diferencias ortográficas y artículos de Rosario/12.

Distancia entre medios según términos y tópicos utilizados

Un experimento interesante realizado en [Fortuna 2009] ubica a los medios analizados en un espacio bidimensional, para apreciar las diferencias y similitudes según los términos y tópicos utilizados.

Siendo que en este trabajo se analizan sólo tres medios, estos se pueden ubicar en dos dimensiones sin pérdida alguna de información sobre las distancias establecidas entre cada par de medios.

En esta sección se trabaja con los artículos de todo el año 2006. El código utilizado se encuentra en la carpeta `media_map/`.

Se toman tres criterios entonces como distancias entre medios:

1. La efectividad de un clasificador que los distingue. Para esto se entrenó el clasificador que mejor desempeño tuvo en el punto anterior, el SVC lineal, con los artículos de cada par de medios. La distancia entre dos medios es el score F_1 obtenido por el clasificador.

2. Los términos utilizados por cada diario. Para esto se utiliza un vocabulario de tamaño fijo de 5000 términos, tomando como corpus todos los artículos del 2006 de los tres diarios. Una vez calculado el vocabulario se toman todos los artículos de cada diario como un solo artículo, con esto se representa cada diario como un vector de tf-idfs, y se calcula la similitud del coseno entre cada par de vectores. La distancia tomada entre cada par de diarios es luego la inversa de esta similitud.

3. La mezcla de tópicos utilizada por cada diario. Para esto se toman todos los artículos del 2006 de los tres diarios como un corpus y se calculan 500 tópicos utilizando LDA. Con estos tópicos se calcula la mezcla de tópicos de los artículos de cada diario, se representa a cada diario como la media de todos los vectores de mezcla de tópicos, y se calcula la similitud del coseno entre cada par de vectores. La distancia tomada entre cada par de diarios es luego la inversa de esta similitud.

Las distancias entre medios obtenidos mediante cada criterio se muestran en la Tabla 13.

	Según la efectividad del clasificador que los distingue			**Según los términos utilizados**			**Según la mezcla de tópicos utilizada**		
	La Nación	**Clarín**	**Página/12**	**La Nación**	**Clarín**	**Página/12**	**La Nación**	**Clarín**	**Página/12**
La Nación	0	0.715	0.791	0	1.059	1.112	0	1.087	1.160
Clarín	0.715	0	0.812	1.059	0	1.156	1.087	0	1.272
Página/12	0.791	0.812	0	1.112	1.156	0	1.160	1.272	0

Tabla 13: Distancia entre pares de medios según distintos criterios

Para obtener una idea visual de la proximidad entre unos y otros medios, se ubican estos tres medios en un espacio bidimensional. Técnicamente se utiliza MDS para ubicar los puntos, pero sólo por

comodidad ya que como se dijo se pueden ubicar los puntos manteniendo inalteradas las relaciones entre las distancias. El algoritmo de MDS en este caso no introduce ninguna pérdida de información. Las distancias entre medios según los tres criterios obtenidos se muestran en la Figura 39.

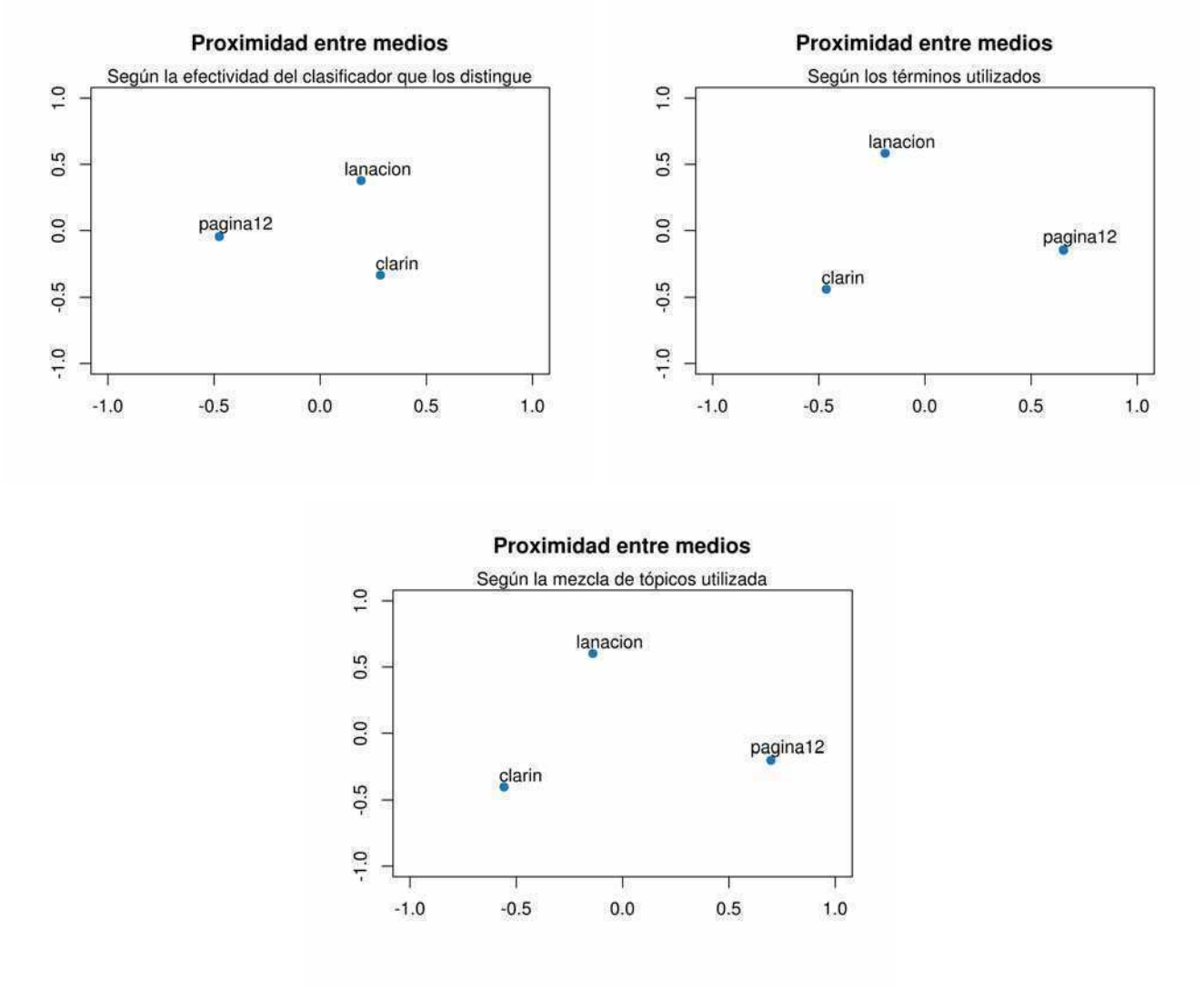

Figura 39: Distancia entre medios según distintos criterios

Como se ve en la Figura 39, los resultados son consistentes de manera independiente al criterio utilizado. Clarín y La Nación se presentan levemente más próximos entre sí, como era de esperarse, y notablemente Página/12 se encuentra en todos los casos un poco más próximo a La Nación que a Clarín.

Mediaticidad de legisladores

Ya buscando resultados que crucen los datos periodísticos con los conjuntos de datos anteriormente analizados, se confecciona un filtro sencillo que decide si un legislador fue mencionado en un artículo periodístico. Un primer intento busca sólo el apellido, pero esta estrategia sufre de excesivos falsos positivos, debido a legisladores que comparten el apellido con otras personalidades mediáticas (Moreno, Sosa, Bielsa), legisladores con apellidos que también son nombres propios de otras cosas (Santander, Merlo, Córdoba), legisladores con apellidos comunes (Pérez, Aguirre) y/o legisladores con apellidos que también son sustantivos comunes que aparecen a veces con mayúsculas por otras razones (Barrios, Puerta, Ríos).

Analizando una muestra de artículos es claro que afortunadamente cuando se menciona a un legislador en una nota periodística generalmente se lo hace por el nombre completo, por lo que se reemplaza el filtro por uno que busca “uno o más nombres seguido de uno o más apellidos del legislador”. La forma “Apellido, Nombre” no se encuentra en ningún artículo del conjunto de datos, por lo que se descarta esta forma.

Se considera armar un filtro más elaborado que considere la posibilidad de que solamente el apellido es suficiente siempre y cuando también se mencione alguna palabra como “diputado” o “legislador”, o bien el bloque al que pertenece el diputado, pero los resultados obtenidos hasta el momento parecen suficientemente buenos.

En la Figura 40 se muestra los primeros 10 legisladores más mediáticos de los años 2006, 2007 y 2008.

2006		2007		2008	
Apellido	**Menciones**	**Apellido**	**Menciones**	**Apellido**	**Menciones**
Macri	1343	Macri	3441	Rossi	788
Carrio	968	Carrio	2301	Sola	704
Rossi	552	Sola	1302	Perez	577
Binner	517	Bullrich	630	Lozano	452
Iglesias	457	Lozano	577	Bullrich	437
Bielsa	411	Binner	510	Pinedo	388
Camaño	340	Perez	492	Kunkel	309
Diaz Bancalari	323	Bielsa	369	Aguad	304
Alvarez	321	Pinedo	333	Macaluse	282
Kunkel	307	Rossi	296	Obeid	241

Figura 40: Legisladores más mediáticos de los años 2006 a 2008

La ausencia repentina tanto de Macri como de Carrió en el 2008 se explica debido a la renuncia del primero en junio del 2007 para asumir como Jefe de Gobierno de la Ciudad de Buenos Aires, y de ella en marzo del 2007 para dedicarse al armado de la Coalición Cívica con vistas a las elecciones presidenciales. Ya las menciones de ambos durante el 2007 inclusive, en su mayoría no son debido al quehacer legislativo sino a sus otras actividades.

Para ver si la atención que reciben los legisladores depende de su postura legislativa, se visualizaron la cantidad de menciones encontrada de cada legislador, en cada medio, por año. La Figura 41 muestra la cantidad de menciones de los legisladores según su punto ideal para el 2007.

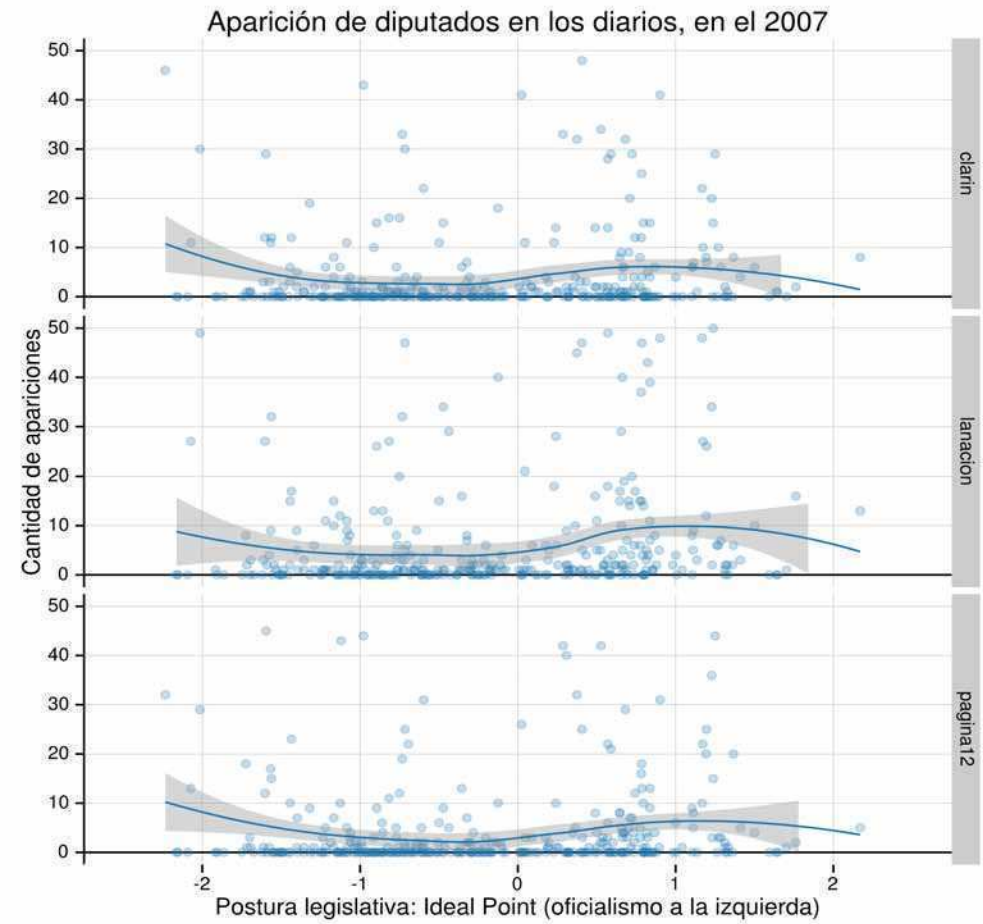

Figura 41: Menciones de diputados durante el 2007

En la Figura 41 se muestra, sobre los puntos, una curva ajustada con loess con una vecindad del 25%, y en gris semitransparente el intervalo de confianza del 95%.

En los tres años analizados parecería que los legisladores con puntos ideales cercanos al 1 (de

oposición no tan extrema) son levemente más mediáticos que el resto, aunque no de manera significativa.

El eje de las ordenadas muestra sólo hasta las 50 menciones, cuando en la Figura 40 es claro que hay varios legisladores con un orden de magnitud más de apariciones en los medios. Estos datos no se muestran debido a que son relativamente pocos, para poder apreciar en detalle lo que sucede con la generalidad de los legisladores, que la mayoría de las veces tiene unas pocas menciones en los diarios. Con sólo mostrar los diputados con hasta 50 menciones se está visualizando más del 90% de los puntos.

Un análisis de interés sería saber si los distintos medios le dan una mayor atención a legisladores de distintas posturas. Las Figuras anteriores no permiten esto con tanta facilidad. Para comparar la cantidad de menciones entre diarios se confeccionaron los diagramas de dispersión con escala logarítmica de la Figura 42. Aquí se utiliza el color para codificar la postura legislativa de los diputados.

Estos datos se muestran para dos años solamente por razones de espacio, pero el comportamiento se repite en el otro año analizado. En principio los resultados son intuitivos: existe una correlación lineal entre las menciones de los legisladores en los distintos medios. Un punto interesante es que nuevamente parece haber una mayor alineación entre las menciones de cada legislador en Clarín y La Nación, que entre Página/12 y aquellos. Efectivamente, si se calcula el coeficiente de determinación R^2 de una regresión lineal entre los tres pares de diarios para estos años, se obtienen los resultados de la Tabla 14.

2006	**2007**	**2008**
lanacion vs. pagina12 R2 = 0.6668199 lanacion vs. clarin R2 = 0.9609833 pagina12 vs. clarin R2 = 0.7155924	lanacion vs. pagina12 R2 = 0.9670197 lanacion vs. clarin R2 = 0.9905516 pagina12 vs. clarin R2 = 0.9624486	lanacion vs. pagina12 R2 = 0.8009839 lanacion vs. clarin R2 = 0.9451747 pagina12 vs. clarin R2 = 0.8564761

Tabla 14: Coeficientes de determinación entre medios para distintos años

Respecto de la postura ideológica, no parece haber un desbalanceo claro en ninguno de los diagramas de dispersión que indique que un medio le esté dando más atención a algunos legisladores que a otros.

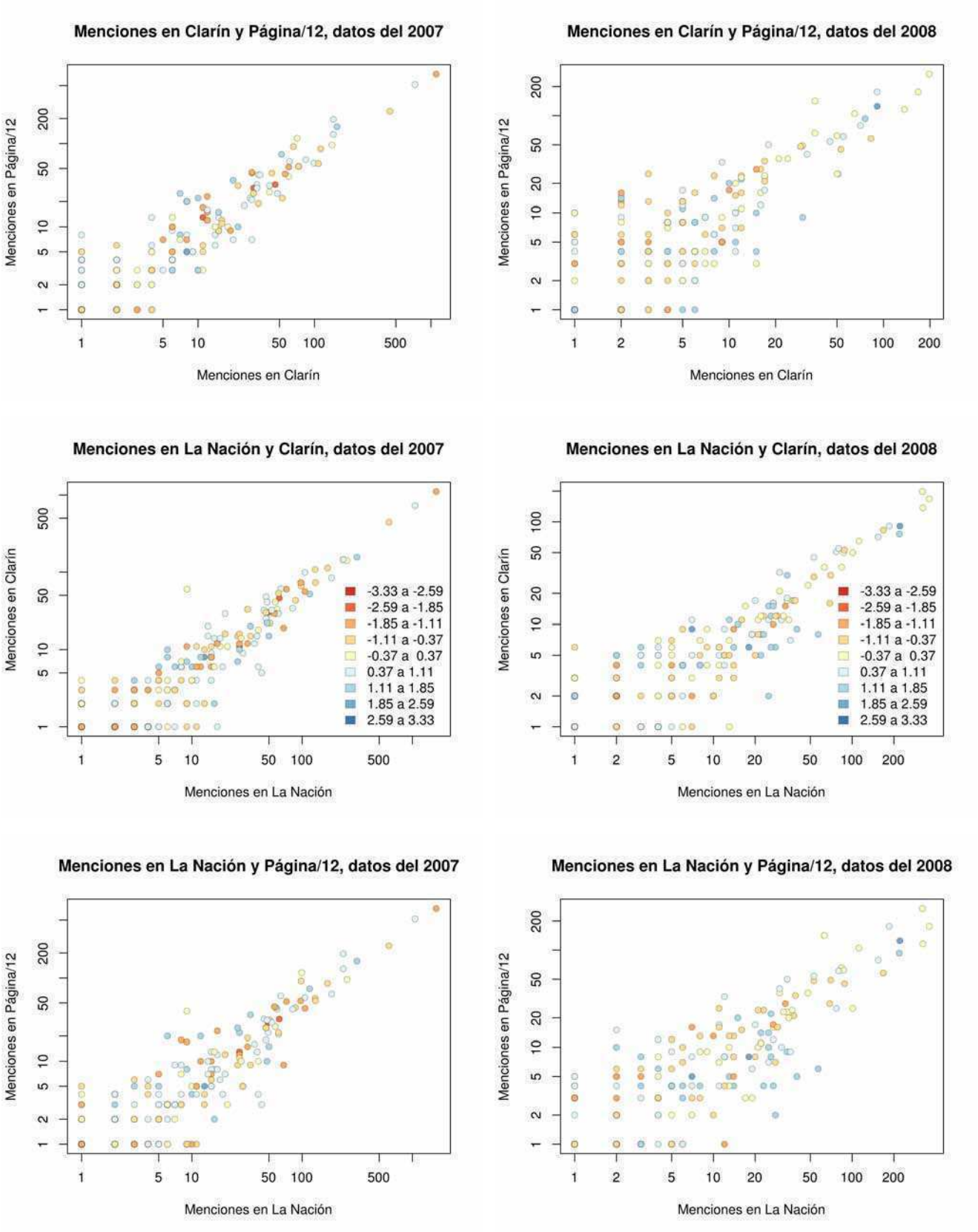

Figura 42: Diagramas de dispersión para datos de 2007 y 2008. El color indica el punto ideal.

Correspondencia entre proyectos y artículos

Contando con un conjunto del orden de un millón de artículos periodísticos y un conjunto de cerca de 30000 proyectos legislativos, se desea decidir cuáles artículos “hablan de” cuáles proyectos.

Aquí “hablan de” necesita tener una definición más específica, ya que no hay muchos artículos que mencionen el número de expediente de un proyecto. Para los fines de este trabajo, se dice que un artículo “habla de” un proyecto si menciona el proyecto mismo, o la temática específica de la que trata el proyecto.

Por ejemplo, dado el proyecto 4706-D-2007, “EXPRESAR REPUDIO A LAS EXPRESIONES DEL REINO UNIDO DE GRAN BRETAÑA CON RESPECTO A LA EXTENSION DE LA ZONA DE EXCLUSION SOBRE LAS ISLAS MALVINAS”, se dice que un artículo periodístico habla de este proyecto si habla de dichas expresiones del Reino Unido, o bien del proyecto de ley, pero no si habla del conflicto de Malvinas más en general.

Este problema se lo puede reducir al siguiente enunciado alternativo: dado un conjunto de artículos periodísticos y **un** proyecto legislativo, decidir cuáles artículos hablan del proyecto dado. Esto lo reduce a un problema de clasificación binaria.

Por ejemplo, si se implementa una solución para este problema con árboles de decisión, para generalizar el resultado a todo el conjunto de proyectos legislativos sería necesario utilizar variables de manera que sean aplicables a otros proyectos. En el ejemplo anterior, “cantidad de palabras del sumario del proyecto mencionadas en el artículo” sería una variable aceptable, pero no “Cantidad de veces que el artículo menciona la palabra Malvinas”. Con esto, estableciendo alguna medida de performance, si se optimiza el clasificador para una muestra arbitraria de proyectos y artículos, debería ser posible extender el resultado inmediatamente a todos los proyectos y todos los artículos.

En particular, si se utilizaran árboles de decisión (o cualquier algoritmo de aprendizaje supervisado), se tendría el problema de generar el conjunto de entrenamiento. Por otro lado, es necesaria alguna métrica de performance. Sin eso, sería difícil encontrar una manera de optimizar el clasificador, o de saber si un cambio realizado en el algoritmo de clasificación resulta en una mejora. Para medir la performance de un clasificador es necesario saber cuáles artículos clasificó “bien”, es decir que es necesario contrastar las predicciones con un conjunto de testeo al que se le ha asignado la clasificación verdadera según un experto de dominio, es decir que han sido clasificados de manera manual.

Se supone entonces que para este problema es deseable utilizar un algoritmo de aprendizaje supervisado, y en algún momento hay que clasificar de manera manual un grupo de artículos para proveer el conjunto de entrenamiento. Tomando esto como un requerimiento, se puede intentar que sea lo más **sencillo** y lo más **significativo** posible.

Para que el proceso de clasificación manual sea **sencillo**, se puede automatizar en gran medida. Evitar tiempos manuales de búsqueda y carga del artículo, proveer un sistema que presente el artículo y el proyecto legislativo en una sola pantalla, que con un click permita asignar una clase a ese artículo respecto del proyecto y pasar al siguiente artículo.

Así se implementa el sistema web de la Figura 43, que presenta un artículo y un proyecto de ley lado a lado, para que con un solo botón se pueda realizar la corrección. El código que implementa este sistema se encuentra en la carpeta `scraper_project/scraper/` como parte de las vistas del sistema Django que se implementó para la adquisición y extracción de los artículos periodísticos y proyectos legislativos.

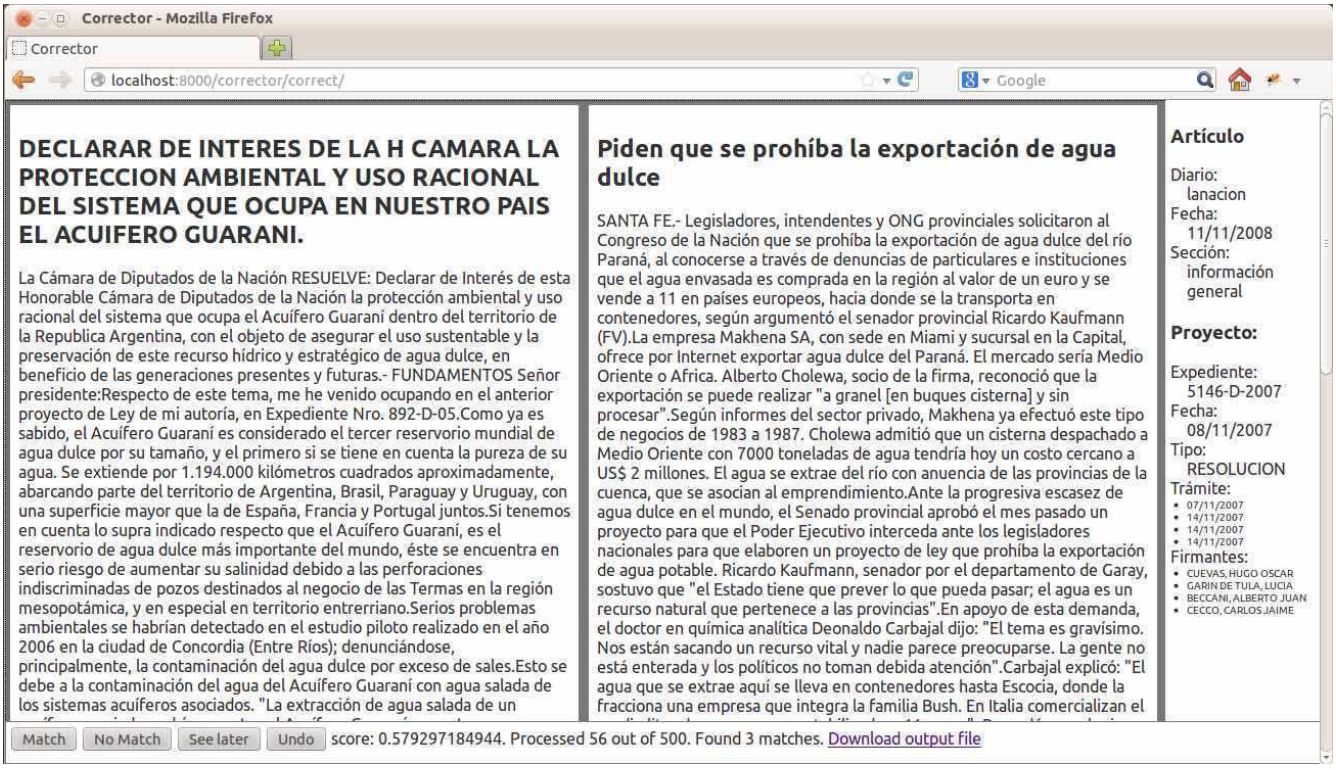

Figura 43: Captura de pantalla del sistema de corrección

Para que el proceso de corrección manual sea **significativo**, es deseable que con la menor cantidad de artículos clasificados manualmente se pueda obtener la mejor noción de la performance del clasificador.

Por un lado, se está trabajando con clases muy desbalanceadas, ya que sólo un puñado de artículos van a hablar efectivamente de un proyecto dado. Esto es fácil de ver ya que cada artículo puede hablar a lo sumo de un proyecto legislativo (o dos, en algunos casos), por lo que necesariamente hay varios miles de proyectos de los que cada artículo no habla. Y esto más allá de que en las secciones de Deportes y Espectáculos no se esperan encontrar muchos artículos que hablen de temas legislativos, cualesquiera sean.

Por otro lado, son justamente los errores de falsos positivos los que más se quieren evitar. La intención del análisis es distinguir el comportamiento de los proyectos mediáticos de los no mediáticos, y las poblaciones se suponen que son de tamaño muy dispar, habiendo más cantidad de proyectos no-mediáticos. Suponiendo esto, si algunos proyectos pasan por menos mediáticos de lo que en realidad son (por errores del tipo falso negativo), con la condición de que no sean la mayoría, no van a modificar el grueso de la población de proyectos no mediáticos como para afectar el comportamiento general. En cambio sí es necesario evitar que proyectos no-mediáticos pasen por mediáticos, ya que incluso un bajo porcentaje de este fenómeno podría afectar las conclusiones de manera irreparable.

Entonces, es necesario encontrar una manera de clasificar manualmente la mayor cantidad de positivos posibles, balanceando así los pesos a priori de las clases.

La manera propuesta de hacer esto es arrancar con clasificadores incompletos que simplemente puntúen la correspondencia entre artículos y proyectos por alguna regla razonable (“artículos que contengan los tres términos de mayor tf-idf del sumario del proyecto” por ejemplo), y luego clasificar manualmente los N artículos que hayan recibido el mayor puntaje. Luego de iterar varias veces, cada nuevo clasificador provee una variable nueva para la posterior clasificación (un score), y permite aumentar el conjunto de entrenamiento con clases más balanceadas.

En cuanto a la performance de estos clasificadores preliminares, ya que no se les impone un punto de corte para asignar una u otra clase, no correspondería confeccionar una matriz de confusión. Pero sí se pueden elaborar curvas ROC con los resultados de la corrección manual, para ver qué tan buen desempeño tienen asignando mayores puntajes adonde corresponda.

Se procede entonces a desarrollar clasificadores preliminares. En todos los casos se trabaja sobre un conjunto de 50 proyectos legislativos tomados al azar de los datos del 2007. A cada clasificador se le piden los 10 artículos con mayor puntaje para cada proyecto, y se clasifican estos de manera manual.

Clasificador 0: Términos del sumario mencionados en el artículo

El primer clasificador que se implementa toma el sumario de todos los proyectos sin lematizar como un corpus, y calcula los tf-idf de los términos. Para cada artículo calcula un score por proyecto que es la suma de los puntajes de cada palabra del artículo, donde el puntaje es el tf-idf de esa palabra en el sumario del proyecto. Inspeccionando los resultados manualmente, tiene un problema severo al favorecer artículos largos.

De los 500 pares proyecto-artículo inspeccionados, se obtienen 54 verdaderos positivos. Se confecciona la curva ROC de este clasificador y se lo muestra en la Figura 44.

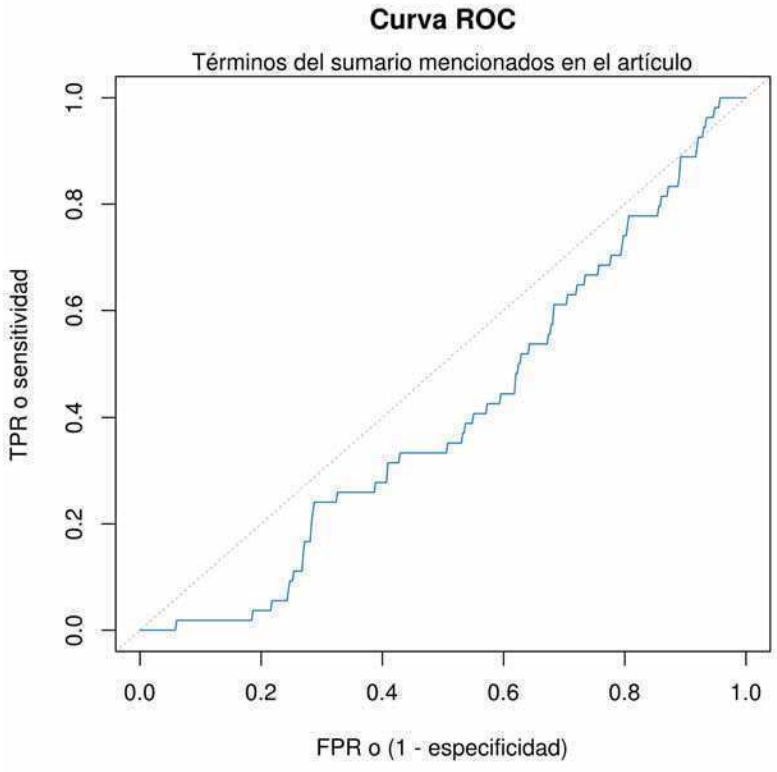

Figura 44: Curva ROC para los términos del sumario mencionados en el artículo

Como puede apreciarse en la curva ROC, por más que el clasificador encuentra 54 verdaderos positivos, su desempeño como clasificador en esta porción de datos es lamentable, de hecho el área bajo la curva ROC es de apenas AUC=0.4075739.

El código de este clasificador se encuentra en la carpeta `proyectos_mencionados/clasificador0/`.

Clasificador 1: Similitud entre tópicos del proyecto y artículo

Se entrena un modelo LDA de 10 tópicos con los cuerpos de todos los proyectos legislativos, tomando tf-idf sobre un vocabulario fijo de 5000 palabras después de lematizar. Utilizando ese modelo se encuentra la mezcla de tópicos de cada artículo, y luego el score asignado a cada par proyecto-artículo es el coseno del ángulo entre los vectores de los tópicos de cada uno.

Con este clasificador se encuentran apenas 4 verdaderos positivos entre los 500 pares mejor puntuados, por lo que se lo descarta.

El código de este clasificador se encuentra en la carpeta `proyectos_mencionados/clasificador1/`.

Clasificador 1.1: Similitud entre tópicos de proyecto y artículo, con 256 tópicos

Considerando al LDA como una forma de reducir la dimensionalidad de los documentos de un espacio con una dimensión por término a un espacio con una dimensión por tópico, puede que el bajo rendimiento del clasificador 1 sea debido a que se tomaron demasiados pocos tópicos, y se perdió la mayoría de la variabilidad del sistema.

Se hace un nuevo intento entonces utilizando 256 tópicos, todavía con un vocabulario de tamaño fijo de 5000 términos después de lematizar. Con esta variante se encuentran 21 verdaderos positivos, y la curva ROC se muestra en la Figura 45.

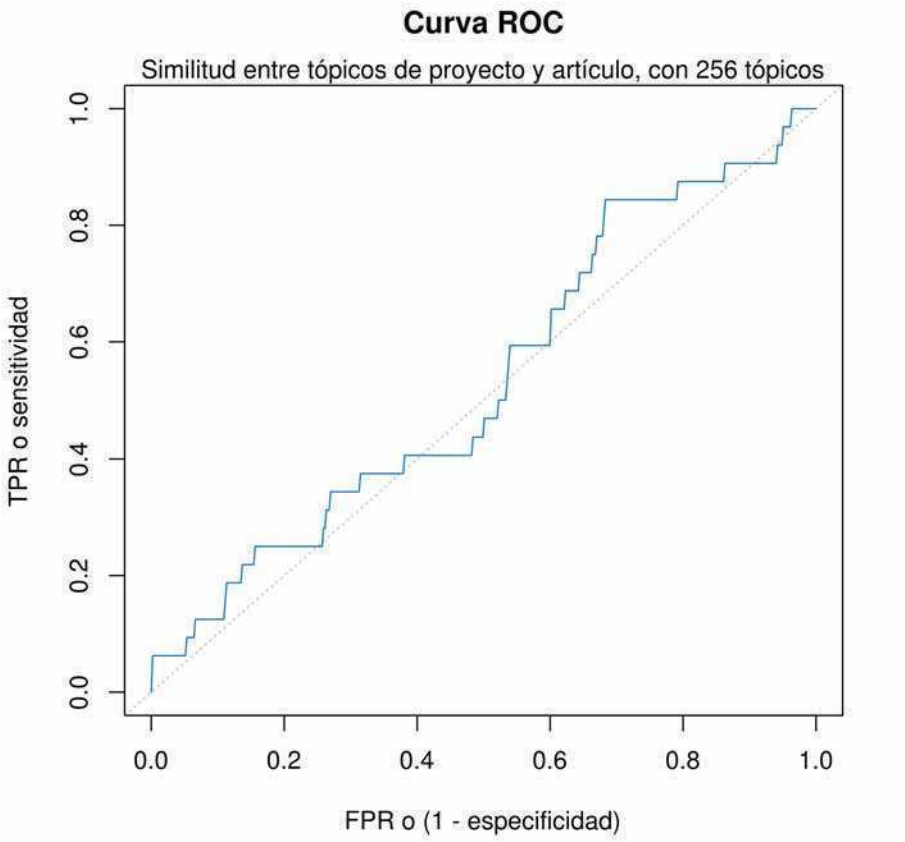

Figura 45: Curva ROC para la similitud entre tópicos de proyecto y artículo

Aunque mejor que antes, el rendimiento todavía no es bueno, y la curva ROC tiene un área bajo la curva de apenas 0.534462.

El código de este clasificador está en la carpeta `proyectos_mencionados/clasificador1.1/`.

Clasificador 2: Similitud entre los cuerpos de proyecto y artículo

Considerando como un corpus el cuerpo lematizado de todos los proyectos, se toma un vocabulario fijo de los 5000 términos de mayor tf-idf global. Cada proyecto y cada artículo luego se representa como un vector con un elemento por cada término del vocabulario, y cada elemento es el tf-idf del término correspondiente en ese documento. El score para un par proyecto-artículo es el coseno del ángulo entre sendos vectores.

Utilizando este clasificador se encuentran 129 verdaderos positivos entre los 500 pares mejor puntuados, y la curva ROC se muestra en la Figura 46.

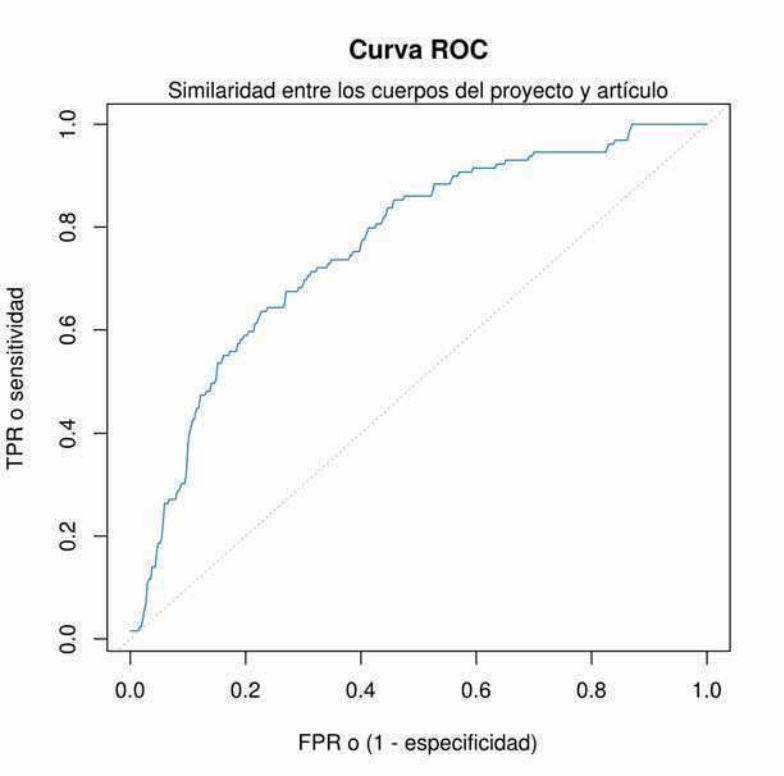

Figura 46: Curva ROC para la similitud entre cuerpos de proyecto y artículo

Este clasificador tiene un comportamiento bastante interesante. El área bajo la curva ROC es de AUC=0.761586.

El código de este clasificador se encuentra en la carpeta `proyectos_mencionados/clasificador2/`.

Clasificador 3: Similitud entre sumario del proyecto y cuerpo del artículo, sin lematizar.

Este clasificador compara, nuevamente por similitud del coseno, el sumario del proyecto con el cuerpo del artículo. Tomando como un corpus el sumario de todos los proyectos sin lematizar, se genera un vocabulario de tamaño fijo de 5000 palabras. Los artículos son reducidos al mismo vocabulario, y se representan todos los documentos como vectores de tf-idfs.

Utilizando esta técnica se encuentran 45 verdaderos positivos, y la curva ROC es la que se muestra en la Figura 47.

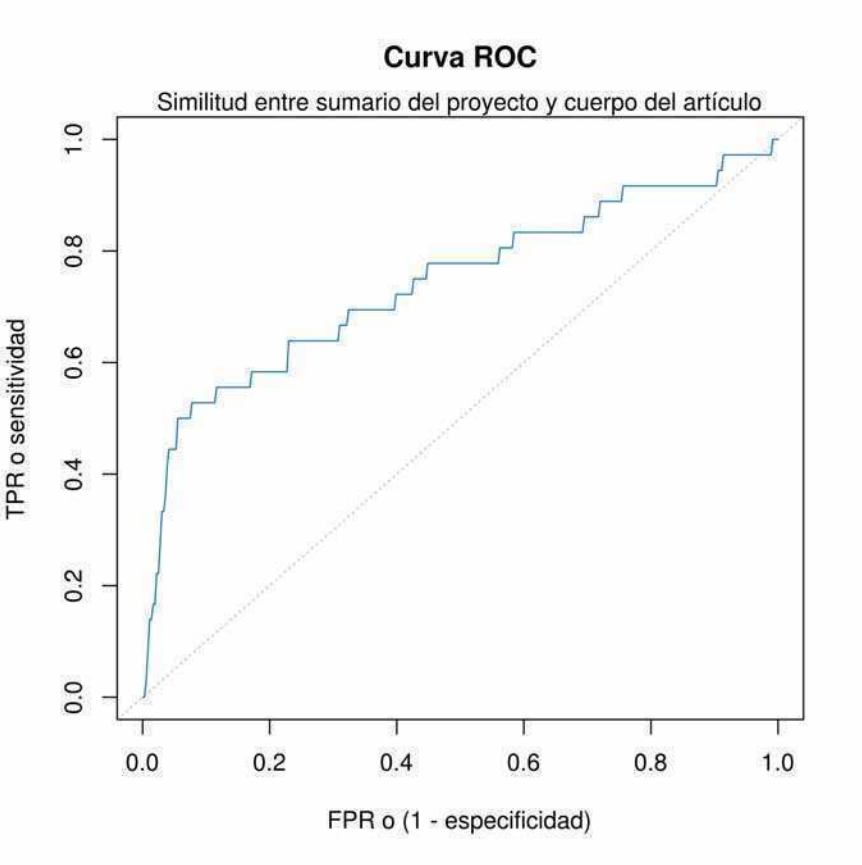

Figura 47: Curva ROC para la similitud entre sumario del proyecto y cuerpo del artículo

Si bien este clasificador no tiene un rendimiento general tan bueno como el clasificador 2, presenta un excelente comportamiento en la primera parte de la curva, para scores elevados es un clasificador bastante confiable. El área bajo la curva ROC es de AUC=0.7416128.

El código de este clasificador se encuentra en la carpeta `proyectos_mencionados/clasificador3/`.

Ensamble con árbol de decisiones

Hasta aquí se lleva generado un conjunto de entrenamiento de unos 2500 pares proyecto-artículo, referidos a 50 proyectos legislativos elegidos al azar del 2007.

Con estos datos se procede a entrenar un árbol de decisiones extensible a todos los proyectos. Para esto se utiliza la librería `rpart` de R.

Las variables que se utilizan para entrenar el árbol inicialmente son las siguientes:

- **Score:** El score según el clasificador 2, que presentó los mejores resultados preliminares.
- **MinDelta:** La mínima distancia en días desde la publicación del artículo a algún tratamiento de la ley (presentación, discusión, aprobación, etc.).
- **Delta:** La diferencia (con signo) en días desde la publicación del artículo a la presentación del proyecto.
- **AbsDelta:** La diferencia (sin signo) en días desde la publicación del artículo a la presentación del proyecto.
- **Sección:** Sección del diario en el que salió publicado el artículo.
- **Tipo:** Tipo del proyecto (Declaración, Resolución o Ley).

El árbol obtenido se muestra en la Figura 48.

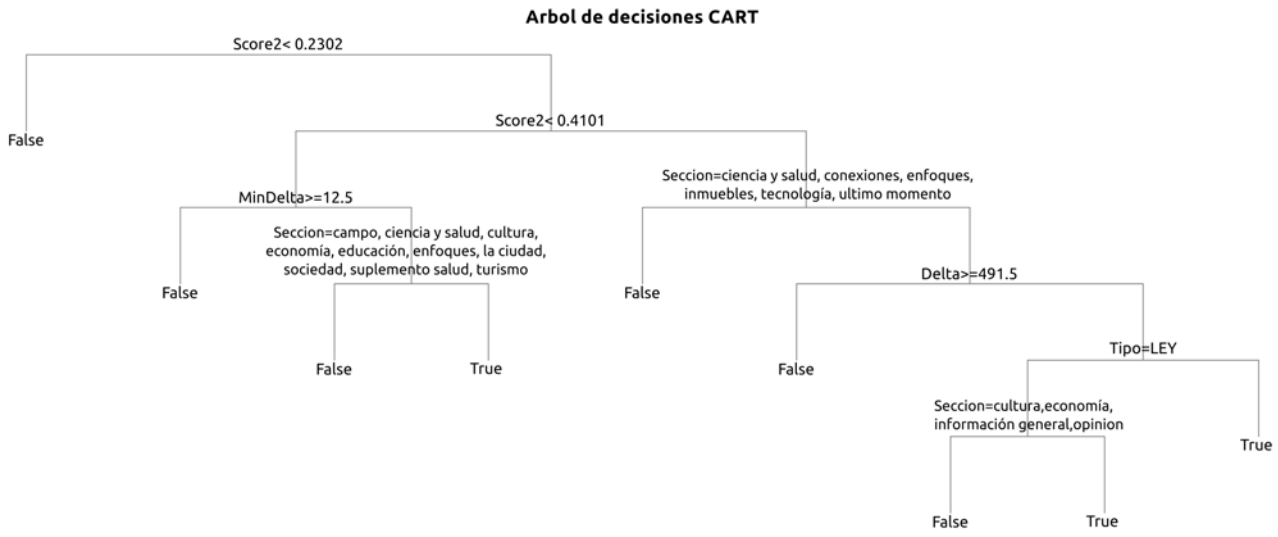

Figura 48: Primer árbol de decisiones entrenado

Para interpretar la Figura 48, cabe mencionar que si la condición de un nodo interno se cumple, se toma la rama de la izquierda.

Con este árbol se obtienen los siguientes valores de performance:

```
          real
predicho  False  True
   False   2409   104
   True      42   125
```

La exactitud (accuracy) en general del clasificador es muy buena, pero esto es debido a las clases desbalanceadas:

$$accuracy = \frac{2409+125}{2409+104+42+125} = 0.9455$$

Tomando otras métricas se obtienen medidas más realistas:

$$sensibilidad\,(recall\,,TPR) = \frac{125}{125+104} = 0.5459$$

$$especificidad\,(TNR) = \frac{2409}{2409+42} = 0.9829$$

$$precisión = \frac{125}{42+125} = 0.7485$$

$$score\,F1 = 2\frac{0.7485 \cdot 0.5459}{0.7485+0.5459} = 0.6313$$

Tanto lo bajo de la sensibilidad como lo alto de la especificidad se explica por lo abundante de la clase negativa.

Como se mencionó con anterioridad, para el análisis que se desea realizar, la métrica más útil es la precisión: si se mantienen bajos los falsos positivos de debería poder mirar diferencias entre el comportamiento de proyectos mediáticos y no mediáticos, aunque algunos proyectos parezcan menos mediáticos de lo que son realmente. La abundancia de proyectos verdaderamente no mediáticos hace que estos proyectos mal clasificados no afecten el comportamiento general.

Para mejorar el desempeño del clasificador se agregan las siguientes variables:

- **Score0:** El score asignado por el clasificador 0.
- **Score0norm:** El score asignado por el clasificador 0 dividido la longitud del artículo. Esta variable intenta compensar el problema de este clasificador por favorecer artículos largos.

- **Site:** El nombre del diario que publicó el artículo.
- **MencionaLegisladores:** 1 si el artículo menciona el nombre de algún legislador, 0 si no.
- **MencionaFirmantes:** 1 si el artículo menciona alguno de los firmantes del proyecto, 0 si no.
- **Score3:** El score asignado por el clasificador 3.

Con estas variables adicionales, el árbol generado es el que se muestra en la Figura 49.

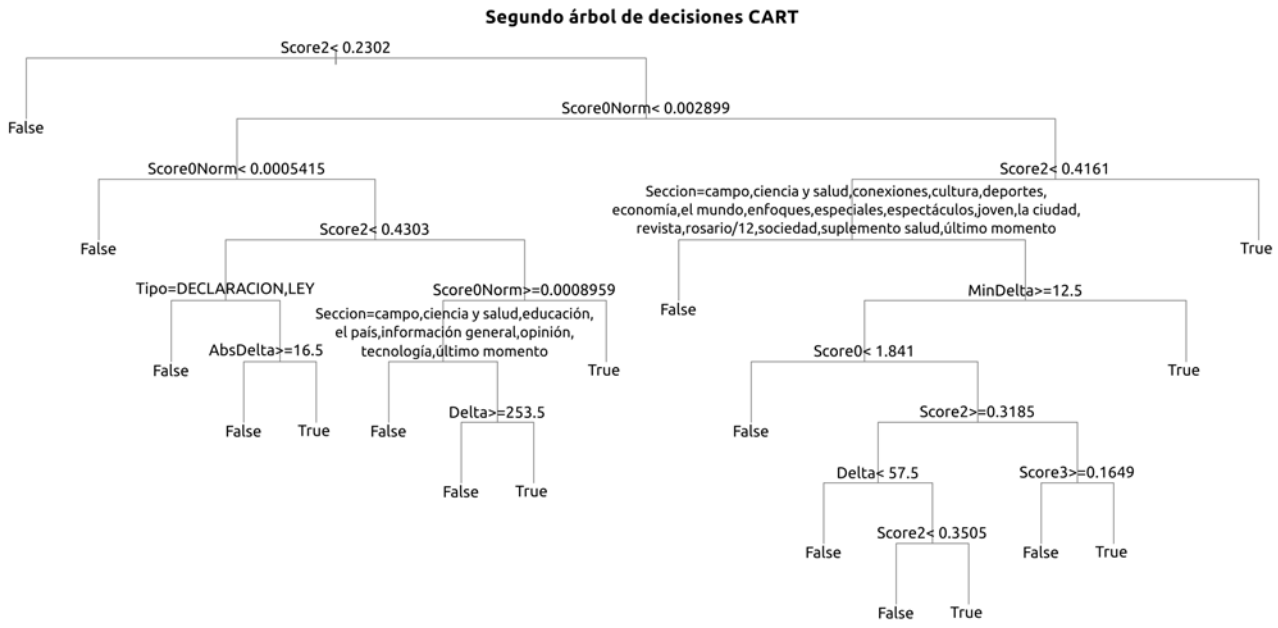

Figura 49: Segundo árbol de decisiones entrenado

Es notable que en ningún momento se utilizan las variables **MencionaLegisladores** o **MencionaFirmantes**. En el caso de **MencionaFirmantes** puede que se deba a que en el conjunto de entrenamiento no hay suficientes artículos que mencionen a los firmantes del proyecto, de hecho hay sólo 13 de estos casos; pero hay más de 450 casos en que el artículo menciona a algún legislador, por lo que debe ser que simplemente esta variable no tiene suficiente relación con la clase buscada.

También es interesante que no se utiliza la variable **Site**, por lo que aparentemente el medio que publica un artículo no es tan relevante para decidir si habla o no de un proyecto legislativo.

Utilizando este árbol se obtiene la siguiente matriz de confusión, evaluando el desempeño con el mismo conjunto de entrenamiento:

```
         real
predicho False True
   False  2430   82
   True     21  147
```

De donde se obtienen las siguientes medidas de performance:

$$exactitud = \frac{2430+147}{2430+82+21+147} = 0.9616$$

$$sensibilidad = \frac{147}{147+82} = 0.6419$$

$$especificidad = \frac{2430}{2430+21} = 0.9914$$

$$precisión = \frac{147}{21+147} = 0.875$$

$$score\,F1 = 2\frac{0.875 \cdot 0.6419}{0.875+0.6419} = 0.7406$$

Ahora, el árbol obtenido posiblemente resulte en overfitting del conjunto de entrenamiento, por lo que es necesario considerar el error obtenido durante la validación cruzada. La librería `rpart` aplica validación cruzada de 10 cruces por defecto, que se considera razonable para este caso.

El pruning del árbol se lleva a cabo mediante el parámetro de complejidad '`cp`', que define qué tanto debe mejorar el ajuste general del modelo una división para llevarse a cabo.

El error relativo en validación cruzada se puede graficar utilizando la función `plotcp`, que se muestra en la Figura 50.

Como se observa, el mínimo error en validación cruzada se obtiene para un valor de cp = 0.012, por lo que se ajusta el árbol a este valor. El árbol final resultante se muestra en la Figura 51.

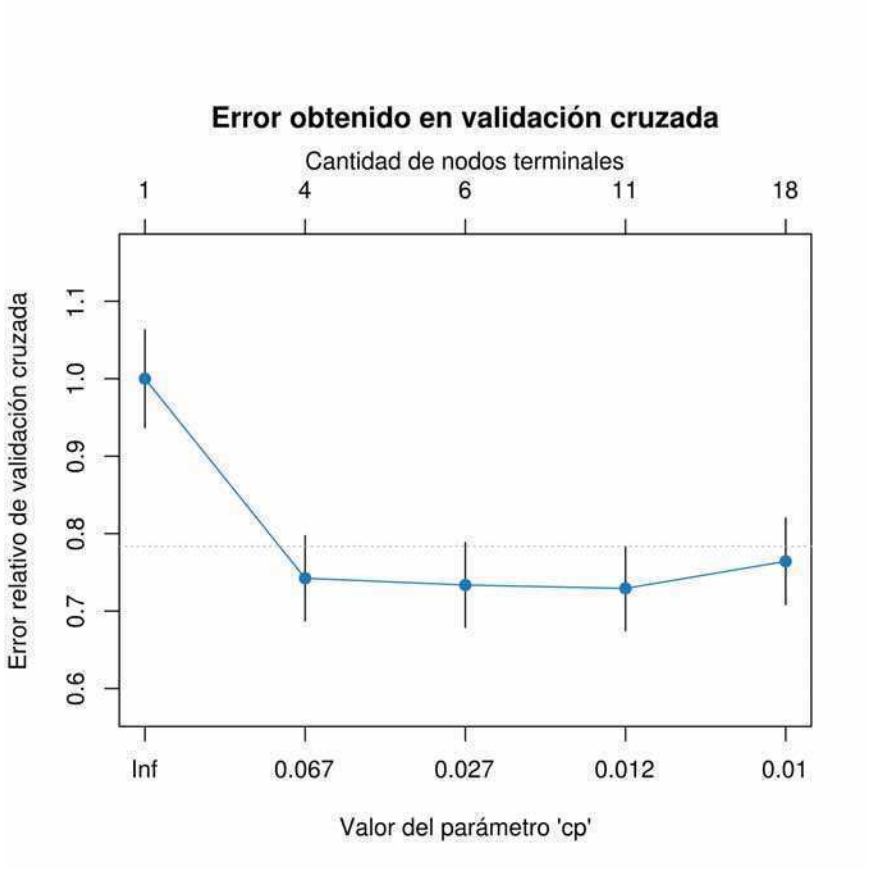

Figura 50: Error relativo obtenido en la validación cruzada

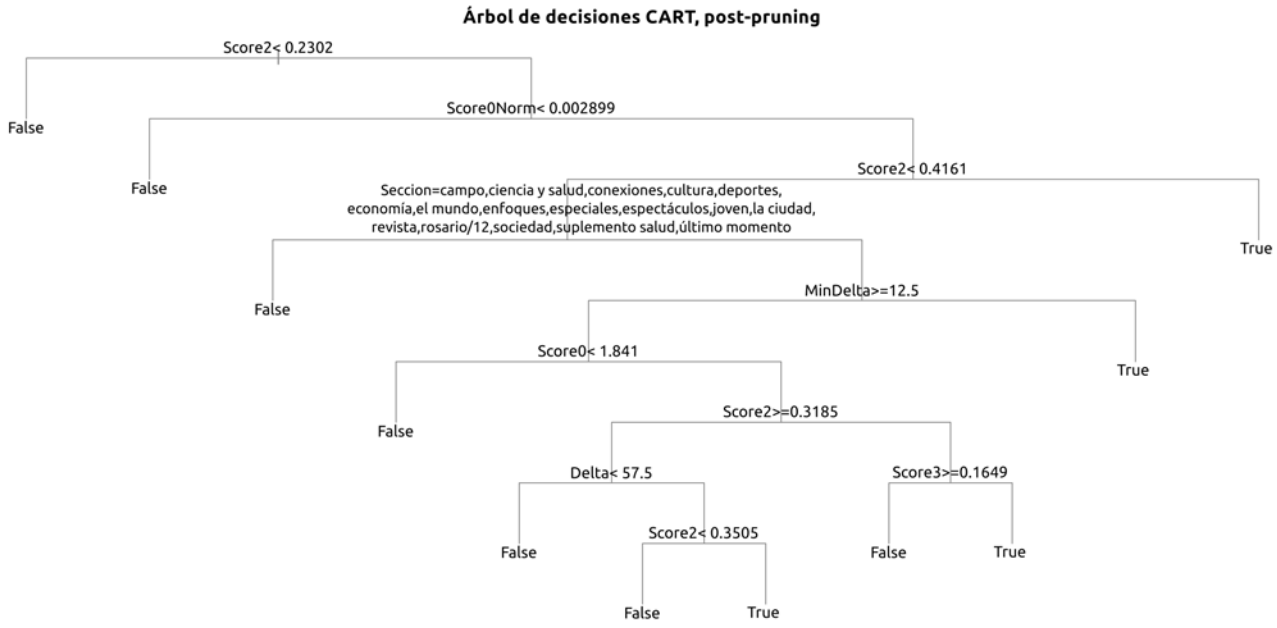

Figura 51: Árbol de decisiones final, con cp=0.012

Con este árbol la matriz de confusión obtenida es

```
predicho False True
   False  2433  106
   True     18  123
```

y las medidas de performance calculadas son:

$$exactitud = \frac{2433+123}{2433+123+106+18} = 0.9537$$

$$sensibilidad = \frac{123}{123+106} = 0.5371$$

$$especificidad = \frac{2433}{2433+18} = 0.9927$$

$$precision = \frac{123}{123+18} = 0.8723$$

$$score\,F1 = 2\frac{0.8723 \cdot 0.5371}{0.8723+0.5371} = 0.6649$$

Aunque la sensibilidad no es mejor que la del árbol inicial, por un lado esto es con el tamaño del árbol ajustado al mínimo error con validación cruzada, y por otro la precisión, que es la métrica que más interesa al presente trabajo, es más de 10 puntos porcentuales más alto.

Utilizando este árbol de decisiones, entonces, se predicen los artículos que hablan de proyectos legislativos para todos los datos obtenidos. En la siguiente sección, ya sobre el final del trabajo, se analizan estos resultados.

Análisis de los resultados predichos

Con los resultados del punto anterior se puede graficar la cantidad de artículos publicados en relación a proyectos legislativos, por fecha. Esto se muestra en la Figura 52.

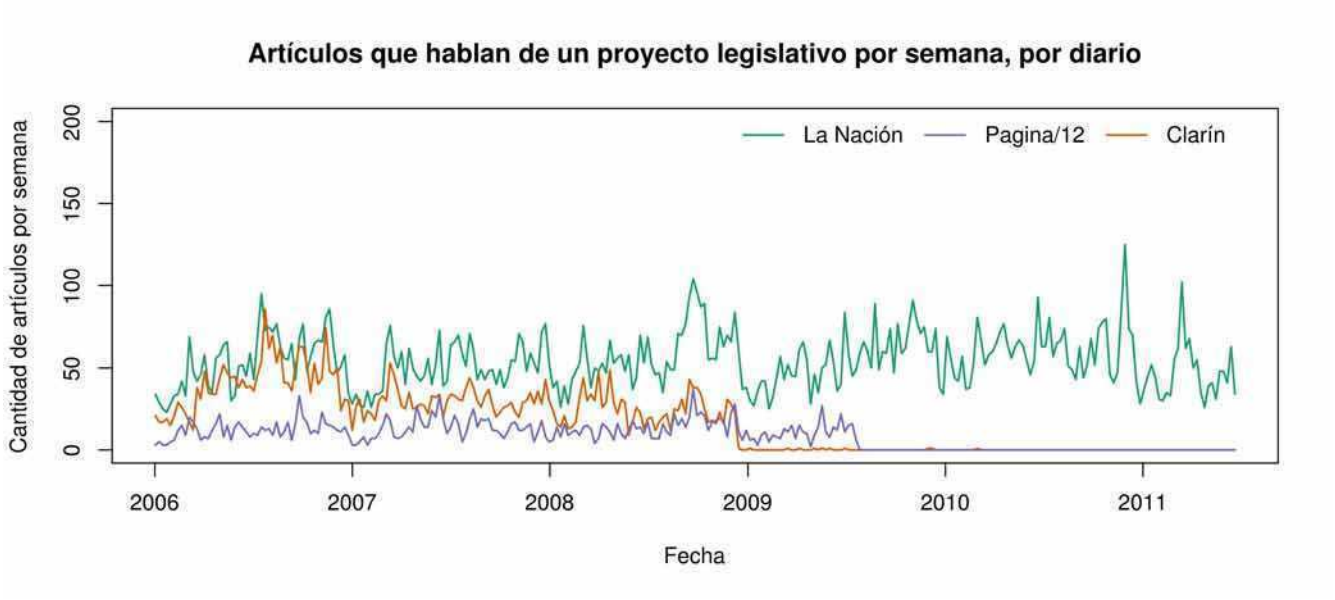

Figura 52: Artículos que hablan de proyectos legislativos, por semana, por diario.

Es notable que para el diario Clarín el número de artículos predichos cae repentinamente para los años 2009 – 2011, y el diario Página/12 los resultados caen abruptamente a mediados del 2009. Esto más allá del hecho de que se recaudaron muy pocos artículos en general para Clarín desde mediados del 2010 en adelante, parece indicar que el clasificador no se ajusta bien a los datos de esos años. Esto probablemente se debe a que el conjunto de datos utilizado para entrenar contiene exclusivamente proyectos del 2007. Sería interesante sin embargo saber qué cambió, en la estructura de los proyectos legislativos o bien de los artículos periodísticos, para que esto suceda.

Sabiendo esto, se procede a analizar los resultados para los años 2006 – 2008 inclusive solamente.

Artículos publicados sobre proyectos de oficialismo y oposición

Uno de los interrogantes planteados al comenzar el trabajo era si los proyectos más mediáticos tenían más o menos posibilidad de ser aprobados. Esto se muestra en la Figura 53 en la columna de la izquierda, para los distintos años. Para esto se separan los proyectos en deciles, según la cantidad de artículos publicados al respecto.

Aunque la variación no es tan grande, se observa que los artículos más mediáticos tienen menor probabilidad de ser aprobados en general.

En la columna de la derecha de la misma Figura 53 se muestra el punto ideal promedio de los proyectos, según su mediaticidad. A mayor cantidad de artículos publicados se nota un aumento en el punto ideal promedio del proyecto. Esto es consistente con los resultados de la Figura 31, que para estos tres años, los proyectos con menor probabilidad de ser aprobados son los de punto ideal más opositor.

Superpuestos a los datos en cada gráfico se traza la recta correspondiente a una regresión lineal para cada año. La misma se construye utilizando la función `lm` de R, y se muestran los valores de la pendiente y su significancia en la Tabla 15.

	Año	Pendiente estimada	Error estándar	T-valor	Prob[>|t|]
Punto ideal promedio	2006	0.031418	0.004478	7.017	0.000111
	2007	0.035669	0.008331	4.282	0.00268
	2008	0.036809	0.007054	5.218	0.000804
Tasa de aprobación	2006	-0.002836	0.001630	-1.741	0.12
	2007	-0.002185	0.002413	-0.905	0.392
	2008	-0.005168	0.001950	-2.65	0.0292

Tabla 15: Pendiente estimada del punto ideal y tasa de aprobación, según la mediaticidad del proyecto

Efectivamente en la Tabla 15 se ve que los valores de variación de la tasa de aprobación no son significativos, pero el punto ideal de los proyectos varía significativamente con la cantidad de artículos publicados al respecto.

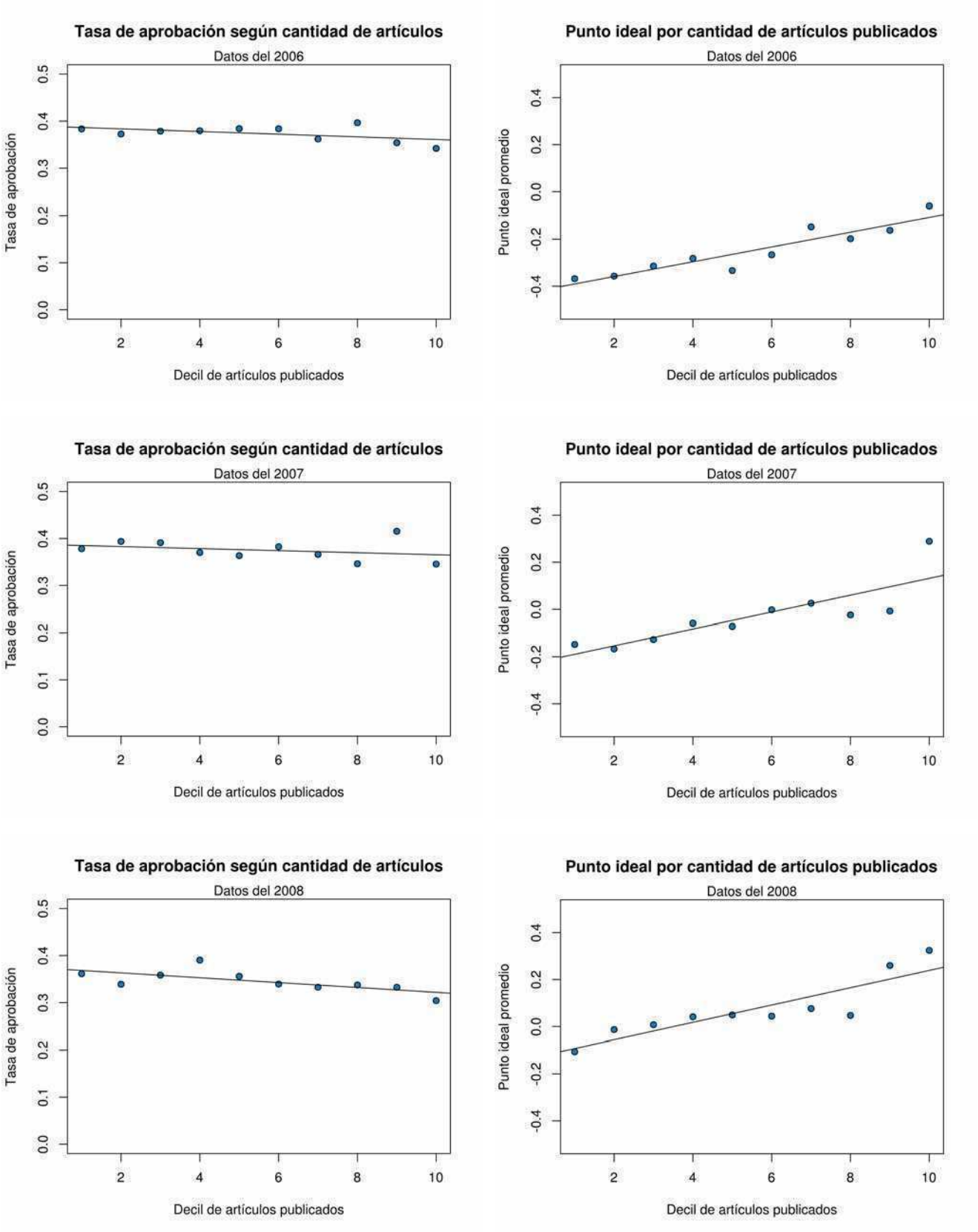

Figura 53: Análisis de proyectos según la cantidad de artículos publicados al respecto

Agregándole una dimensión más de interés al análisis, en la Figura 54 se desdoblan estos gráficos según el medio en el que se publican.

Allí se observa que Clarín y La Nación consistentemente todos los años le dan más atención a proyectos de una postura legislativa más opositora. Página/12 en cambio, y sobre todo en el 2008, mantiene constante el punto ideal promedio en el caso de los proyectos más mediáticos. De la misma manera, mientras que los proyectos más mencionados por Clarín y La Nación ven una menor tasa de aprobación, los proyectos más mediáticos para Página/12 ven una disminución en el 2006, un aumento en el 2008, y en el 2007 mantienen una tasa de aprobación aproximadamente igual a los proyectos menos mediáticos.

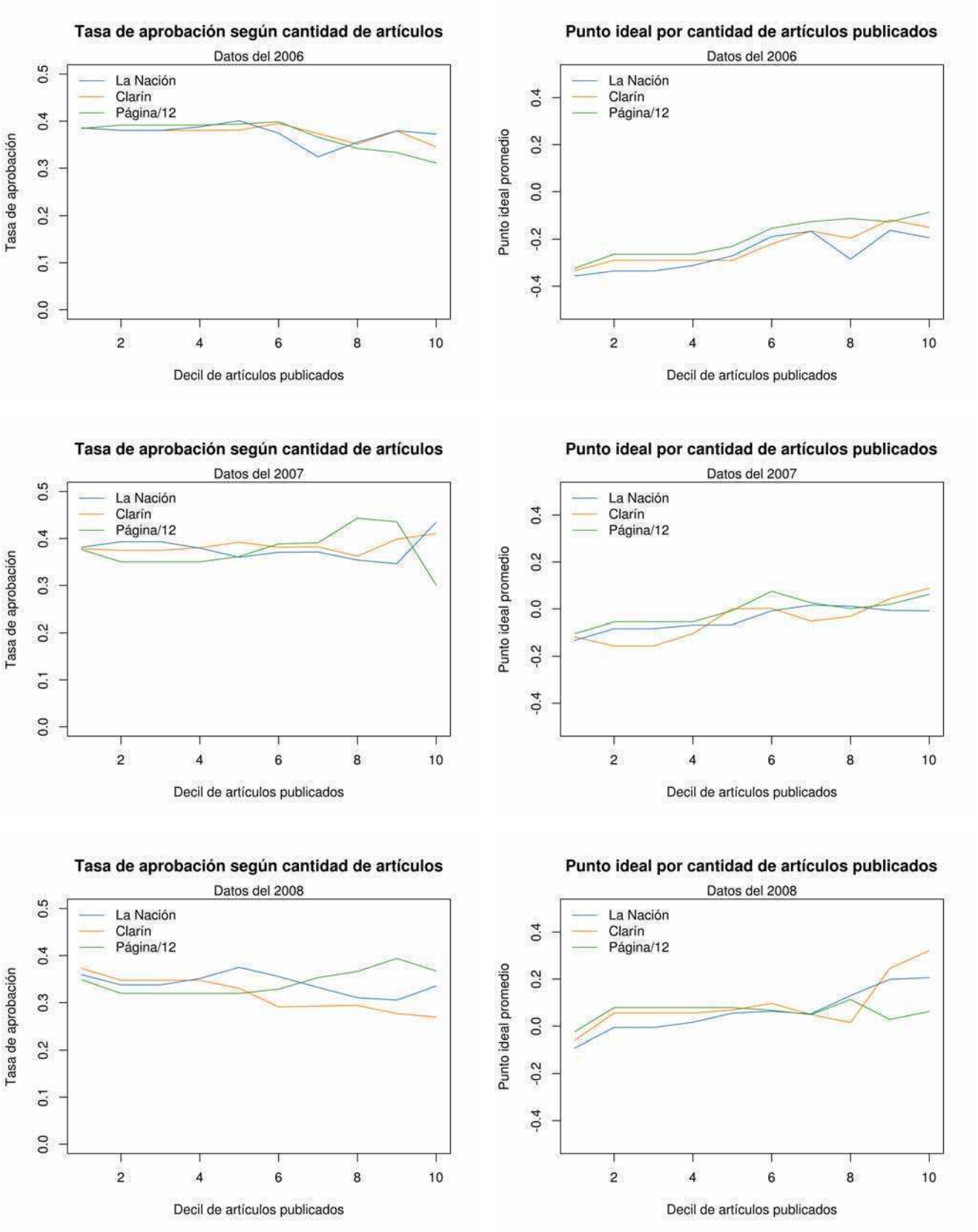

Figura 54: Análisis de proyectos según la cantidad de artículos publicados, por diario

Conclusiones

En este trabajo se exploraron algunas de las técnicas disponibles para analizar los datos legislativos y mediáticos disponibles en la Argentina.

Como se mostró a lo largo del trabajo, los puntos ideales de los legisladores nos dan un "índice de oficialismo" extremadamente útil para posteriores análisis, ya que no es necesario informarlo respecto de la filiación política autodeclarada por los legisladores. En este trabajo se combinaron sólo con un par de conjuntos de datos adicionales, pero existen variadas fuentes de datos adicionales que podrían explorarse utilizando estos puntos ideales. Sin ir más lejos que el mismo sitio de la Cámara de Diputados, se encuentran disponibles también los discursos de los legisladores dentro de la cámara, y todos los debates de las sesiones, las asistencias y conformación de las distintas comisiones, que podrían explorarse utilizando estos puntos ideales.

En cuanto a las menciones de legisladores y proyectos en los artículos resultó interesante, aunque quizás intuitivo, encontrar la variación de la cantidad de artículos publicados según el punto ideal y tasa de aprobación del proyecto en la Figura 53. Habiendo dicho eso, los años analizados son dentro de todo un período acotado de tiempo, y un contexto político bastante similar (limitado al Kirchnerismo, con mayoría – o casi mayoría – del oficialismo en diputados). Sería interesante realizar este análisis con datos de contextos históricos distintos, para ver el comportamiento de los distintos bloques en distintas situaciones. Durante los años analizados, los medios parecen tener un comportamiento en algún sentido similar al de los legisladores, como se vio en la sección de *Comparación de posturas legislativas: Recalde, Macaluse y Quiroz*: si bien los medios que están alineados presentan estadísticas más similares entre sí, que dos medios tengan posturas distintas no implica que las mediciones observadas sean significativamente distintas.

Resulta además interesante que consistentemente los medios le brindan más atención a los proyectos opositores, y que estos a su vez son los de menor tasa de aprobación. Si bien se encontró mayor similitud entre La Nación y Clarín como se esperaba, al iniciar el trabajo también se esperaba encontrar diferencias más marcadas entre el comportamiento de los medios que se consideran cotidianamente "opositores" y "oficialistas". Esto puede ser o bien porque la intuición está equivocada y la realidad es más compleja que los medios oficialistas por un lado y los opositores por el otro, o bien que las herramientas utilizadas en este trabajo no son las ideales para detectar estas diferencias. Notablemente no se analizó la valencia de los artículos respecto de los legisladores y los proyectos; un trabajo de

minería de opiniones y análisis de sentimientos posiblemente brinde resultados más marcados.

El conjunto de datos de los proyectos legislativos presentados brindó resultados interesantes e inesperados, en cuanto a la variación de la tasa de aprobación según la conformación de la cámara, como muestra la Figura 31. Posteriores análisis sobre este dato, también con datos de otros años sería interesante para ver cómo ha variado a lo largo de los años el enfrentamiento de los diputados dentro de la cámara baja.

Finalmente, sería interesante poder implementar algunos de los resultados obtenidos en este trabajo en un sistema de monitoreo legislativo, disponible para el público en general, que mantenga actualizados los resultados a medida que avanza el período legislativo. Incluso los resultados que resultaron de interés tangencial en esta tesis, como las técnicas de agrupamiento y segmentación de los legisladores o la medición de la entropía de cada votación podría resultar de suma utilidad al momento de elegir candidatos para la próxima elección legislativa.

Apéndice A – Versiones de dependencias utilizadas

Para todo el código Python adjunto se utilizó la versión 2.7.3 disponible en http://www.python.org/. La siguiente es una lista de los paquetes de terceros utilizados, con su respectiva versión.

- BeautifulSoup 3.2.0 http://www.crummy.com/software/BeautifulSoup/
- Django 1.4 https://www.djangoproject.com/
- PIL 1.1.7 http://www.pythonware.com/products/pil/
- Argparse 1.2.1 http://code.google.com/p/argparse/
- Gensim 0.8.4 http://radimrehurek.com/gensim/
- Httplib2 0.7.4 http://code.google.com/p/httplib2/
- Matplotlib 1.1.1rc http://matplotlib.org/
- Mercurial 2.0.2 http://mercurial.selenic.com/
- Numpy 1.6.1 http://www.numpy.org/
- Pdfminer 20110515 http://www.unixuser.org/~euske/python/pdfminer/
- Psycopg2 2.4.5 http://initd.org/psycopg/
- Pydot 1.0.2 http://code.google.com/p/pydot/
- Pygooglechart 0.3.0 http://pygooglechart.slowchop.com/
- Reportlab 2.5 http://www.reportlab.com/
- Rpy2 2.3.2 http://rpy.sourceforge.net/
- Scikit-learn 0.11 http://scikit-learn.org/
- Scipy 0.9.0 http://www.scipy.org/
- Virtualenv 1.7.1.2 http://www.virtualenv.org/

Para el código R adjunto se utilizó la versión 2.14.1 disponible de http://www.r-project.org/ con los siguientes paquetes y librerías:

- RColorBrewer 1.0-5 http://cran.r-project.org/web/packages/RColorBrewer/
- Ggplot2 0.9.3 http://had.co.nz/ggplot2/
- Rpart 4.1-0 http://cran.r-project.org/web/packages/rpart/
- Wordcloud 2.2 http://cran.r-project.org/web/packages/wordcloud/
- SparkTable 0.9.6 http://cran.r-project.org/web/packages/sparkTable/
- Pscl 1.04.4 http://cran.r-project.org/web/packages/pscl/
- Mixtools 0.4.6 http://cran.r-project.org/web/packages/mixtools/

Apéndice B – Lista de palabras eliminadas

A continuación se incluye una lista de las palabras que se eliminaron del análisis por considerarse que aportan poco valor semántico (stopwords). Se debe tener en cuenta que se en todos los casos se filtran los stopwords luego del paso de lematización, por lo que no es necesario incluir ninguna forma flexionada.

además	crear	explicar	mejor	poder	tanto
ahora	cuando	forma	menos	poner	te
algo	dar	gente	mes	por	tema
alguno	de	gran	mi	porque	tener
allí	deber	hablar	mientras	que	tiempo
ante	decir	hay	mismo	qué	tipo
antes_de	dejar	hoy	mucho	quedar	tm_d
año	desde	hs	mundo	querer	tm_min
ar	después	ir	muy	quien	todo
argentino	después_de	haber	nada	realizar	tomar
argentina	diario	hablar	ni	saber	tras
así	día	hasta	no	salir	tratar
aunque	donde	hacer	nombre	sam	uno
ayer	durante	la	nos	se	último
bajo	el	las	nt	seguir	uu
bien	él	le	nuevo	según	varios
bueno	ellos	llamar	nuestro	ser	venir
cada	en	llegar	nunca	si	ver
casi	encontrar	llevar	otro	sí	vez
com	entonces	ln_m	página	sino	vida
como	entre	lo	país	siempre	vivir
cómo	es	los	para	sin	www
con	estar	luego	parecer	sobre	ya
conocer	ese	más	parte	sólo	yo
contar	este	mayor	pasar	su	
contra	estudio	me	pero	también	
cosa	ex	medio	poco	tan	

Referencias

[Alemán 2009] Eduardo Alemán, Ernesto Calvo, Mark P. Jones, y Noah Kaplan: “Comparing Cosponsorship and Roll-Call Ideal Points,” Legislative Studies Quarterly, No. 34, p.87 – 116, 2009.

[Benaglia 2009] Tatiana Benaglia, Didier Chauveau, David R. Hunter, y Derek Young: “mixtools: An R Package for Analyzing Finite Mixture Models”. Journal of Statistical Software, 32(6), p. 1 – 29, 2009.

[Blei 2009] David M. Blei y John D. Laffery: “Topic models”. En Ashok Srivastava y Mehran Sahami (eds.) “Text mining – Classification, clustering and applications” p. 71 – 93. Ed. Chapman & Hall/CRC, 2009.

[Blei 2011] David M. Blei: “Introduction to Probabilistic Topic Models” Communications of the ACM, 2011

[Canes-Wrone 2002] Canes-Wrone, Brandice, David W. Brady, John F. Cogan: “Out of Step, Out of Office: Electoral Accountability and House Members’ Voting.” American Political Science Review 96: 127–40., 2002.

[Clinton 2004] Joshua Clinton, Simon Jackman, Douglas Rivers: “The Statistical Analysis of Roll Call Data” American Political Science Review Vol. 98, No. 2, Mayo 2004

[Conway 2012] Drew Conway y John Myles White : “Machine Learning for hackers”, Ed. O'Reilly, 2012

[Cox 2001] Trevor F. Cox y Michael A. A. Cox: “Multidimensional Scaling”, 2ª Edición, Ed. Chapman & Hall/CRC.

[Everitt 2011] Brian S. Everitt, Sabine Landau, Morven Leese, y Daniel Stahl: “Cluster Analysis” 5ª Edición. Serie Wiley de Probabilidad y Estadística, Ed. Wiley, 2011.

[Fan 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, y Chih-Jen Lin: “LIBLINEAR: A library for large linear classification”. En Journal of Machine Learning Research N°9, 2008.

[Fortuna 2009] Blaz Fortuna, Carolina Galleguillos, y Nello Cristianini : “Detection of Bias in Media Outlets with Statistical Learning Methods ”. En Ashok Srivastava y Mehran Sahami (eds.) “Text mining – Classification, clustering and applications” p. 26 – 50. Ed. Chapman & Hall/CRC, 2009.

[Honrado 2000] Asunción Honrado, Ruben Leon, Ruairi O'Donnel y Duncan Sinclair: “A Word Stemming Algorithm for the Spanish Language”, Ponencias de la Seventh International Symposium on String Processing and Information Retrieval, p. 139 – 145, 2000.

[Martin 2002] Andrew D Martin, Kevin M Quinn: “Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999”, Journal de Análisis Político Vol. 10, No. 2, p. 134-153 Oxford University Press, 2002.

[Padró 2010] Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes y Irene Castellón: “FreeLing 2.1: Five Years of Open-Source Language Processing Tools”, Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), ELRA. La Valletta, Malta, 2010.

[Rennie 2003] Jason D. Rennie, Lawrence Shih, Jaime Teevan, y David Karger: “Tackling the poor assumptions of naive bayes text classifiers.” en Machine Learning – International Workshop then Conference – Vol. 20, N°2, p. 616, 2003.

[Rojas 1996] Raul Rojas: “Neural Networks: A Systematic Introduction”. Ed. Springer, 1996.

[Rosenthal 2004] Howard Rosenthal y Erik Voeten: “Analyzing roll calls with perfect spatial voting: France 1946–1958”. En American Journal of Political Science, Wiley Online Library, 2004.

[Sirovich 2003] Lawrence Sirovich: “A pattern analysis of the second Rehnquist U.S. Supreme Court ” Laboratorio de Metemática Aplicada, Departamento de Ciencias Biomatemáticas, Mount Sinai School of Medicine, Nueva York, 2003.

[Wiemer-Hastings 2004] Peter Wiemer-Hastings: “Latent Semantic Analysis” DePaul University School of Computer Science, Telecommunications, and Information Systems , Chicago, Noviembre 10, 2004.

[Wright 2002] Wright, Gerald C., Brian F. Schaffner. “The Influence of Party: Evidence from the State Legislatures.” American Political Science Review 96: 367–79. 2002.